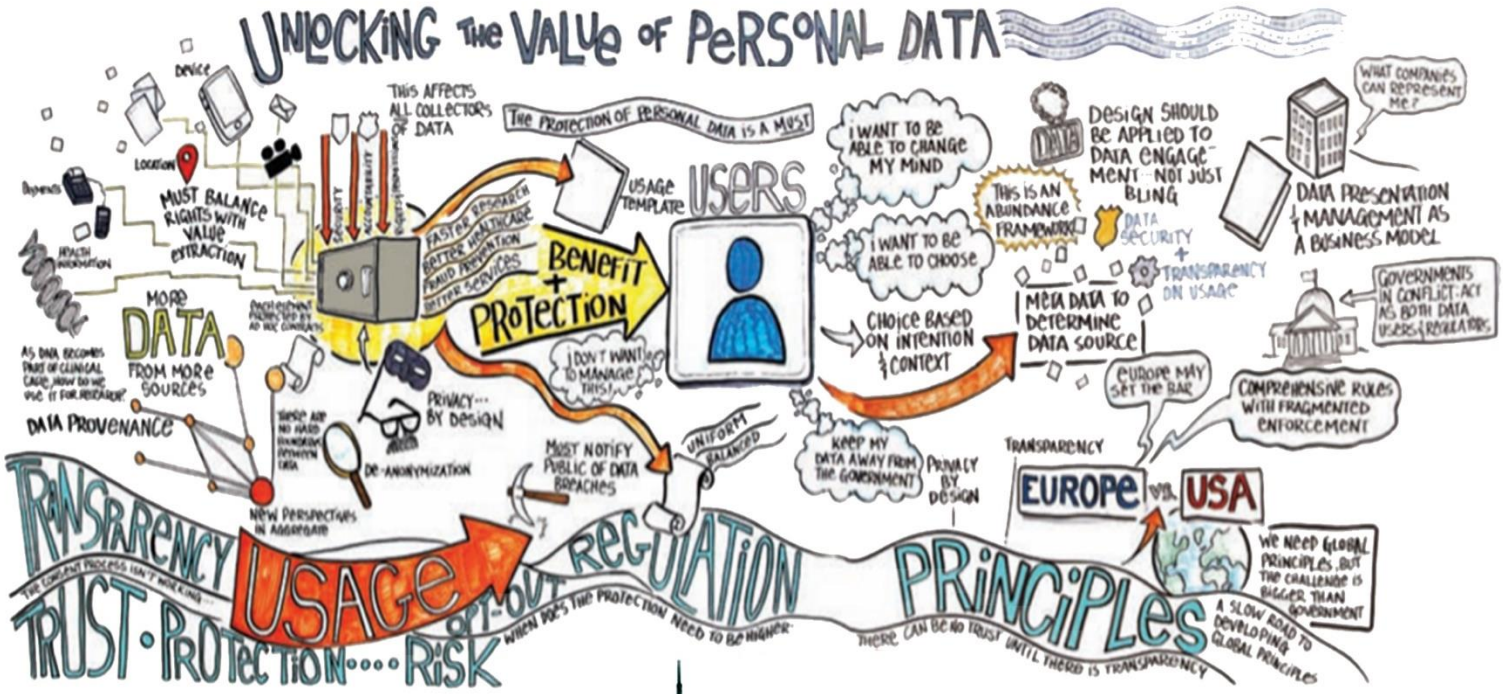
enabling **data marketplaces** with **multi-party computation** (mpc)

an exploratory study investigating the implication of the maturation of multi-party computation technology to the architecture and the threat landscape of the data marketplaces



JEEVAN KUMAR

master thesis

TU Delft

Cover by ***Suhasini Sharana Munoli***

Cover Images from ***thevalueweb.org*** and ***pixels.com***

Enabling Data Marketplaces With Multi-Party Computation (MPC)

An Exploratory Study investigating the Implication of the Maturation of Multi-Party Computation (MPC) technology to the Architecture and the Threat Landscape of the Data Marketplaces

A **Master Thesis** submitted to **Delft University of Technology**
in partial fulfilment of the requirements for the degree of

Master of Science

in

Management of Technology

Faculty of Technology, Policy and Management,

by

Jeevan Kumar

Student Number: ***4743822***

To be defended publicly on ***September 26, 2019*** at ***11:00 AM***.

Graduation Committee:

Chair:	***Dr. ir. G.A. (Mark) de Reuver*** Section of Information and Communication Technology
First Supervisor:	***Dr. -Ing. T. (Tobias) Fiebig*** Section of Information and Communication Technology
Second Supervisor:	***Dr. ir. C. (Carlos) Hernandez Ganan*** Section of Organization and Governance

This thesis is confidential and cannot be made public until September 26, 2019.

An electronic version of this thesis is available at http://repository.tudelft.nl/

Abstract

The emergence of the Data Marketplaces is the latest iteration in the phenomenon of data-driven transformation of the world. Data marketplaces have emerged as a new form of data-driven business models which enable trading of data between the data owners/providers and data consumers by providing the necessary technological and non-technological infrastructure. These features present an alternative to the cumbersome logistics currently involved in searching, buying and selling data; thus, simplify the data supply chains between the data-driven business entities. However, they suffer to take off into mainstream success because of a myriad of reasons. Of all the reasons, 2 of them are focused in this thesis. Firstly, the difficulty involved in architecturally enabling a data marketplace platform as the prospective enabling technologies are still immature. Secondly, the uncertainty associated with the commodification of data which comprises of the intellectual property enforcement of data (data ownership), privacy and confidentiality breach (threats), regulatory ignorance (implication of GDPR), reluctance of businesses from participating because of the previous reasons et cetera. This reason is collectively referred as due to the uncertainty around the threat landscape of the data marketplaces. Multi-Party Computation (MPC) technology provide a solution to these problems. Through its capabilities to preserve the confidentiality of data architecturally and thereby securing the interests of the data actors with respect to the uncertainty of the threat landscape around data, MPC can enable safe and secure data sharing between data actors. This characteristic of MPC can help data marketplaces to overcome their challenges and foster their realisation. However, since MPC cannot handle the scale of real-life application, it is not mature enough yet to be incorporated into real-life data marketplaces. An EU funded project called SafeDEED: Safe Data-Enabled Economic Development, proposes to overcome the scalability issue and intends to achieve the maturation of MPC for real-life application. Building upon this forecast, a research was conducted to investigate the implication of the maturation of MPC technology towards the 2 problems faced by data marketplaces, architectural and threat landscape; and the same is documented in this thesis.

The research was performed through the development of 4 conceptual models. The first 2 models comprised of Pre-MPC Data Marketplace Platform (a high-level architecture of the data marketplace platform) and Post-MPC Data Marketplace Platform (MPC incorporated architecture). The difference of these 2 models explicated the implication of MPC on the architectural aspects of the data marketplaces. The second 2 models comprised of Pre-MPC Threat Model (threats to data marketplaces prior to MPC incorporation) and Post-MPC Threat Model (threats after MPC incorporation). The difference of these 2 models explicated the implication of MPC on the threat landscape of the data marketplaces. Both the differences were summed up to obtain 2 new conceptual models and subsequent hypotheses which collectively constitute the deliverable of the thesis.

The development of the former 4 conceptual models was carried out in 2 phases. Firstly, the *Conceptualisation* phase where the first iteration of models was developed using desk research methods. Secondly, the *Validation phase* where the models were subjected to validation through expert interviews. The validation was carried out through qualitative data analysis on the lines of Middle-Ground Approach of theory generation with the first iteration of the models serving as initial specification. As a result, the second iterations of the 4 conceptual models were generated which represented more valid conceptualisations. These contributed towards developing the theoretical framework and its subsequent conceptual models and hypotheses which reflect the implication of the maturation of MPC technology to the architecture and the threat landscape of the data marketplaces. In this way, the thesis presents a business application for MPC technology in data marketplaces once the former becomes mature; thus, potentially contributing towards the realisation of data marketplaces and thereby, fostering the data-driven economy in Europe.

"DATA IS THE NEW OIL"

- *Almost every Organization in the World right now*

Acknowledgements

They say the process of scientific research not only betters the way we understand the world but the way we understand ourselves. I say amen to that as I emerge a more well-rounded person concluding this challenging yet massively-enriching master's thesis.

For an introvert from the south of India living in the most comfortable bubble one can have in the likes of a well-paying job, cool colleagues, great friends, fun weekends planned well in-advance, awfully nurturing family and an amazing city of Bengaluru, it surely was not a popular move to get out in pursuit of exploring the world and one's self with it. But with one shot and one opportunity, I made the brave call (rather thoughtless) to come to the Netherlands despite constantly feeling my palms sweaty, knees weak and arms heavy about the decision I made. But I chose wisely with TU Delft as life here so far has turned out to be a roller coaster ride with personal bests and absolute worsts. It was all worth it because, in the end, it did matter as I believe to have grown into the best version of my intellectual, professional, social, physical, mental, emotional and even surprisingly, culinary self. Quite a number of interesting people were instrumental in this journey of mine and this long account of appreciation is to make these very people uncomfortable with my genuine gratitude.

Firstly, my Graduation Committee. Dr Mark de Reuver, thank you for your insightful and straightforward assistance, be it during the courses that you taught me as well as during my thesis. Your detailed replies (which is every master student's dream) helped significantly in shaping my research as the replies precisely delivered the right insight to fuel the decision making in solving my most crucial concerns. I would also like to appreciate your timely effort towards connecting me to the experts who contributed to my research without which I would have been hung out to dry. You did all these with just one face-to-face talk, 4 email chains and 2 feedback documents; and that's why I was fortunate to have you as the Chair of my committee. Dr Tobias Fiebig. I would give you most of the credit for my conversion from a student to a researcher. I thank you for encouraging me to explore whatever I wanted to do as part of this research rather than forcing myself into the confines of an already formulated research. I take a deep sense of pride in the fact that I was able to execute my first research project independently; right from the genesis to its fruition, and I give your patient and open-ended supervision all the credit. Having you as my First Supervisor is the reason for the research acumen, I have developed so far over the past 8 months. And finally, Dr Carlos Ganan. Having you as my Second Supervisor helped me to get a cybersecurity point of view towards the thesis. I thank you for bringing your insights to the table which was very different from that of the rest in the Committee.

Secondly, the support system in my social sphere. My legit F.R.I.E.N.D.S circle of PANDAS from India. Though I came away from you people, your presence never seized to exist (painfully); thanks to the innovation of phone call over WhatsApp and group call over Messenger. Thank you, Sindhu, Topi, Shreya, Kuri, Abhi, Sammi, Chandu, Meena, Kulli, Sudha, Praveen for being the audience and laughing at my every joke. It is all because of you my sense of humour has grown which helped me survive in this far off land. Thank you, Geeks of Wasseypur (Saurabh, Lakshya and Kiran) for keeping the geek in me alive even remotely. Thank Zhihao, Nidhi and Paul for hosting the innocent me, right from my first day in the Netherlands and for being the best flatmates anyone could wish for. Thank you, Sushmitha for being such a pleasurable trouble in life. Thank you Maruthi Ram Kashyap for being the only fellow traveller on my side throughout this journey of Masters. Your friendship is one of the best things about my chapter in the Netherlands. Thank you, Nihal for re-establishing the long-lost pizza connect from India all the way in the Netherlands. Thank you, Concordia and the indoor cricket Gully Boys for literally providing me with the playground in which I could blow my steam off and also, reinforce the sportsperson in me. Thank You, Delftse Dance Crew for resurrecting the dying dancer in me and providing me the space to nurture my sense of humour. Thank you, Srushti, Akash and Narayani for letting me infiltrate your chemical gang and

making me a part of everything you do even though I am just a spirit-uplifting outsider. You guys were a great part in making my life in TU Delft a memorable one and I am super glad that we became friends so fast which made up for all the time we lost in the first year.

It is incomplete if I do not thank my fictional friends who gave the most necessary motivation for me to get through the toughest parts of the last 2 years. Being a pop culture fan myself, I am glad I got to experience the rightest movies and tv shows relevant to my state of mind during these 2 years. Thank You, Ben Wyatt, Leslie Knope, Ron Swanson, Tom Hoverford, Donna Meagle and Chris Traeger for being the source of inspiration in the moments when I was lost during my literature study. Thank You, Michael Scott, Dwight Schrute, Jim Halpert, Pam Beasley and Andy Bernard for inspiring to be your true self no matter what the situation forces you to be. Valery Legasov, Boris Shcherbina and Ulana Khomyuk for so beautifully teaching me to do the right thing irrespective of the scale of the consequences. Thank you, Jake Peralta, Raymond Holt, Amy Santiago, Terry Jeffords, Rosa Diaz and Doug Judy for portraying not to take the work you do too seriously for your own good. Thank You, Eleanor, Chidi, Tahani, Jason, Janet and Michael for questioning what it means to be a human. Thank you, Jon Favreau and Andrew Rea for making the art of cooking exciting for a mortal like me. Thank You Hasan Minhaj, Aziz Ansari, Dave Chappelle, Bill Burr and Kunal Kamra for being the comic relief and simultaneously keeping me in touch with the world during the research. And finally, thank you, Ludwig Goransson and Ramin Djawadi for composing the music of Black Panther and Game of Thrones which was playing throughout in the background when I was writing this thesis.

And finally, my Family, Thank You, Mom, for getting accustomed to technology just to keep in touch with me; even though you hate tech. Your video calls which used to last for a maximum of 15 minutes once a week would fuel me enough to get through the week. Thanks for letting me be me, trusting my decision making and being by my side watching full of hope as I figure out my path in life. Thank you, my annoying sister for just being you. I am aware you care even in your annoying little way and thanks for your unconditional support and the phone that you bought me which was a lifesaver here for me. Thank you, Dad, for making adorable attempts to translate the love you have for me. I totally understand that, and I downplay the same way you do. I hope I make you proud of this master's degree.

Last but definitely, not the least, the love of my life, the Leslie Knope for my Ben Wyatt, my permanent roommate, the one who would cheer for my effort even in my failures, the one who would do any means necessary just to see me succeed. Thank you, Suhasini, daughter of Sharana Munoli for being the bundle of joy that you are in my life. Words will fall short to express how instrumental your existence has been in my path of getting to where I am today. I am glad I listened to you to root our lives to come to Europe. It has led to this and I cannot wait to start our lives officially together that we have been planning for 8 years. By the way, Thank You for designing the cover of the thesis with your great artistic instincts.

I could go on, but it is getting to a point of being inappropriate. However, without these people, I would definitely, not have made it. Finally, I would like to thank TU Delft for giving me the space not only to become an educated manager and a researcher but also a stronger and fearless person. The journey of the thesis, however testing it may look and feel at times, the completion of it would be the most fulfilling moment yet in life. However, you will get there only by undergoing the necessarily pain stalking yet the most stimulating experience out of which you will emerge with the mastery over the art of perseverance. With this subjective opinion, this is Jeevan, Signing off!

May the force be with you,

Delft, September 2019

Table of Contents

List of Figures

List of Tables

Abbreviations

ATT&CK	Adversarial Tactics, Techniques & Common Knowledge Framework
B2B	Business-to-Business
CAPEC	Common Attack Pattern Enumeration and Classification
CIA	Confidentiality, Integrity, Availability
COI	Community-of-Interest
CTSA	Cyber Threat Susceptibility Analysis
(D)DoS	(Distributed) Denial of Service
DevOps	Software Development (*Dev*) and Information-Technology Operations (*Ops*)
DL4LD	Data Logistics for Logistics Data
DLT	Distributed Ledger Technology
DMA	Data Market Austria
DMP	Data Marketplace Platform
DREAD	Damage, Reliability, Exploitability, Affected Users, Discoverability
DVT.	Data Valuation Technology
EDM.	Enterprise Data Market
FFIEC.	Federal Financial Institutions Examination Council
FSS	Financial Services Sector
GDPR.	General Data Protection Regulation
HLA.	High-Level Architecture framework
HLCTM.	High-Level Cyber Threat Modelling Framework
HSSEDI.	Homeland Security Systems Engineering & Development Institute
IDDIL/ATC	Identify the assets; Define the attack surface; Decompose the system; Identify attack vectors; List threat actors; Analysis & assessment; Triage; Controls.
Intel's TARA.	Threat Agent Risk Assessment
IoT	Internet of Things
Microsoft SDL	Security Development Lifecycle
MITRE'S TARA	Threat Assessment and Remediation Analysis
MPC.	Multi-Party Computation
NGCI	Next Generation Cyber Infrastructure
NIST	National Institute of Standards and Technology
OWASP	Open Web Application Security Project
PII	Personal Identifiable Information
PPT	Privacy-Preserving Technologies
SaaS	Software as a Service
SafeDEED	Safe Data Enabled Economic Development
SDLC	Systems/Software Development Life Cycle
SESAR	Single European Sky ATM(Air Traffic Management) Research
SSH	Secure SHell encryption protocol
STRIDE.	Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege
TTP	Tactics, Techniques, Procedures

1

Introduction

Over the past decade, data has emerged to be one of the most valuable business resources disrupting and fuelling the transformation of a myriad of industries, thereby justifying the statement, "Data is the new oil" (Hartmann, Zaki, Feldmann, & Neely, 2016). Big Data Revolution has encouraged businesses to adopt data-driven innovation which could potentially improve their productivity and efficiency. The knowledge extracted from the data and in many cases, the data itself have helped organizations to create enormous value in the form of data-driven decision making and data-based products respectively (Davenport, 2006; Brynjolfsson & McAffee, 2012). The value of data economy fostered by the activities of generation, collection, storage, processing, distribution, analysis and exploitation of data; is expected to be around €700 Billion by 2025 in Europe alone (Lupu, 2018). To achieve this forecast, a well-coordinated marriage between demand and supply of the data is necessary. Data Marketplaces play a fundamental role in orchestrating this marriage by offering a platform equipped with different services for data owners to sell their data and for data seekers to find good quality data of their interest (Koutroumpis, Leiponen, & Thomas, 2017; Deichmann, Heineke, Reinbacher, Wee, 2016). Through simplifying data supply chains by overcoming the cumbersome logistics currently involved in searching, buying and selling data, data marketplaces could help in establishing data ecosystems comprising of a network of organizations across different industries, and thereby could boost the data-driven economy.

However, despite their significance, the number of successful commercial data marketplaces are surprisingly low, and the number of failed ones is very high (Koutroumpis et al., 2017). The reason for this phenomenon can be twofold.

- Firstly, an intrinsic reason that it is challenging to design and set up a technologically viable platform to trade data (Koutroumpis et al., 2017). The immaturity and unavailability of enabling technologies presents a difficulty in developing a sound platform for a data marketplace.
- Secondly, an external reason associated with the lack of trust among data actors which is manifested by the uncertainty associated with data security because of the sensitive nature of data (Lupu, 2018). For data-driven innovation to flourish, it is crucial that the data owners share their data. However, several issues discourage them to do so. Some of the issues are: lack of clarity in the implementation of *General Data Protection Regulation* (GDPR), risks associated with privacy violations, threats associated with the business and cyber space around the data marketplaces. Fully concerned with these issues, data actors exhibit reluctance towards participating in data marketplaces. Since the data marketplaces are platforms prone to positive externalities, this reluctance of data actors directly implies a barrier which significantly contributes towards their slow and delayed flourishing of data marketplaces.

Multi-Party Computation (MPC) technology, as claimed by SafeDEED, provides a solution to these issues with their capabilities to safeguard the data and thereby, the interests of the data actors (Lupu, 2018). SafeDEED claims that MPC technology has the potential to enable safe and secure data sharing among the data actors in a confidentiality-preserving and privacy-preserving way; in the sense that, the data actors can share the knowledge existing in their data to the needful consumers without having to share the physical data itself. This proposition is interesting for the data marketplaces as their core business process is to enable secure trading and sharing of data.

This thesis aims to make use of this promise of the MPC technology and investigate if it enables the data marketplaces to overcome their barriers; in turn helping the data marketplaces achieve their true potential of prospering the data market in Europe. With this agenda, this thesis further contributes to the ongoing research of both, the data marketplaces and the MPC technology.

The rest of the chapter is structured as follows. Section 1.1 establishes the broader context which gave way for the research in this thesis. Section 1.2 explains the research problem addressed by this thesis by explicating the problem statement, the knowledge gap, the research objective, the main research question and the complementing research tasks and sub-research questions. Section 1.3 presents the research design of the thesis with the research framework and further, sets up the methodology used during desk research and empirical research. Section 1.4 specifies the scope at which the thesis is operating specifying at its lowest-level. Section 1.5 provides an overview on the contribution of the thesis. Finally, section 1.6 concludes the chapter by providing an illustration of the structure of the thesis report.

1.1 Context of the Research

The research is carried out as part of the SafeDEED*: Safe Data Enabled Economic Development* project (Lupu, 2018). It is a consortium of research organizations from cryptography, data science, business model innovation and legal domains across Europe, which is funded by the *European Union's Horizon 2020 Research and Innovation* program, to foster and accelerate the data-driven economy in Europe. SafeDEED proposes to develop technologies to promote the data sharing culture among the organizations and foster the data-driven economy in Europe. The technologies proposed by SafeDEED are of 2 categories*: Secure Multi-Party Computation* (MPC) and *Data Valuation Technologies* (DVT). With these technologies, SafeDEED aims to encourage the data owners to share their data by taking care of 2 crucial things:

- by enforcing the security aspects of the data sharing (through Secure MPC); and
- by explicating the value of the data held by the data owners to both data owners and data consumers (through DVT)

By ensuring these 2 aspects, SafeDEED aims to incentivise the data actors into indulging in data trading. The research of this thesis focusses only the Multi-Party Computation (MPC) technology and builds on the propositions as conceptualised in the research proposal of SafeDEED project (Lupu, 2018).

1.2 Research Problem

The researchers have proposed Multi-Party Computation (MPC) technology as a crucial enabler of the data marketplaces which has the potential to orchestrate safe and secure data trading (Roman & Stefano, 2016; Lupu, 2018). However, this is just a proposition which has not been investigated yet

owing to the reason that MPC technology has not achieved the desired level of maturation for it to be applied in real-life cases. SafeDEED claims to achieve this desired level of maturation with the help of different research organizations across Europe.

This proposition by SafeDEED about maturing MPC technology can finally unlock the possibility of MPC technology enabling the data marketplaces to the latter's full potential and save them from their delayed rise towards mainstream adoption. This is a researchable agenda where it can be researched how exactly the mature MPC technology enables the data marketplaces now that there is a path laid out by SafeDEED towards the maturation of MPC technology for real-life application. This is the focal research problem of this thesis which is formalised into the following problem statement,

"Data marketplaces suffer heavily from a myriad of barriers which restrict them towards mainstream adoption. Through the research towards maturing MPC technology for real-life application, SafeDEED provides an opportunity for the data marketplaces to overcome their barriers as matured MPC technology, with the ability to handle real-life scale, can potentially enable the data marketplaces to function successfully. As a result, this is a need to investigate the logistics aspect of how exactly the matured MPC technology can enable the data marketplaces. With Europe being in transition towards fully embracing the data-driven philosophy, this is the right time to perform this study so that the data marketplaces can be completely realised and are fully functional by the time, Europe masters the data-driven philosophy"

This thesis contributes towards solving this research problem by focussing only on the 2 problems which affect the successful functioning of data marketplaces (*as introduced earlier*);

- the problem of designing and setting up a technologically viable data marketplace platform (***architectural*** *aspects*).
- the problem related to the uncertainties associated with data sharing and the sensitive nature of data. In this category, we shall specifically focus on the issue of the threats associated with the data marketplaces (***threat landscape***).

Essentially, the research conducted as part of this thesis involves understanding the architectural implications and the implications to the threat landscape of the data marketplaces, thereby explicating the significance of the maturation of MPC technology for the data marketplaces which could potentially be instrumental in the functional realisation of the data marketplaces.

1.2.1 Knowledge Gap

Firstly, related to the *architectural* aspects of the data marketplaces, there has been no investigation of how MPC technology can be incorporated into the data marketplace platform architecturally. Related to this, there exists another problem that there has been no research related to the architectural aspects of data marketplaces. The reasons for this can either be that the architectural information is confidential proprietary information for the real-life data marketplaces to disclose to the research community; or also that the research area of architectural aspects of the data marketplaces is in its infancy and is not explored proactively yet. As a result, there exists no architecture of a data marketplace platform in the literature. So, there is a need to build an architecture which reflects a generic data marketplace platform.

Secondly, it is necessary to investigate the effect of MPC technology on the *threat landscape* of the data marketplaces in order to understand the positive as well as negative implications of MPC technology towards the threats associated with the data marketplaces. Here exists another problem that the threats associated with the data marketplaces has never been identified yet by the research community. So, there is a need to explore the threat landscape of the the data marketplaces to identify the threats that affect them.

1.2.2 Research Objective

The research objective (*RO*) of this thesis which signifies the potential deliverable intended to solve the research problem was formulated as follows,

RO: "To understand the implication of the maturation of Multi-Party Computation (MPC) technology for the architecture and the threat landscape of the Data Marketplaces"

The research objective entails the following steps:

- Firstly, to understand the phenomenon of the data marketplaces and consequently, develop an *architecture* of a generic data marketplace platform.
- Secondly, to explore *threat landscape* of data marketplaces and to identify the threats which affect their functioning.
- Thirdly, to understand how the MPC technology can be *incorporated* into the previously-built architecture and to deduce what does this imply architecturally to the data marketplaces
- Finally, to deduce how the incorporation of MPC technology *affects* the threat landscape of the data marketplaces (*both positive and negative effects*) and what does the same imply to the latter.

Essentially, the research in this thesis is an amalgamation of the 3 subject areas: *Data Marketplaces, Threat Modelling* and *MPC Technology*.

1.2.3 Research Question

An exploratory research question was formulated to reflect the research objective as not much is known about the phenomenon of data marketplaces and not enough theory is available on the application of MPC technology in data marketplaces. The same serves as the main research question of this thesis and is formalised as follows,

RQ: What can be the implication of the maturation of Multi-Party Computation (MPC) technology for the architecture and the threat landscape of the Data Marketplaces?

1.2.4 Research Tasks and Sub-Research Questions

The research was divided into 2 phases: ***Conceptualisation phase*** and ***Validation phase.*** The conceptualisation phase signifies our desk research which was performed using literature study and other desk research methods. Based on the acquired knowledge from the literature and further

desk analysis, theoretical concepts were developed which gave rise to the *artefacts* serving the research objective. The conceptualisation phase was further broken down into 3 research tasks.

- ***RT1:*** *To build an architecture of a generic data marketplace platform.* This involved subtasks which signify, first to figure out a *methodology* for the task and then to *execute* the actual task. The 2 subtasks were,
    - Firstly, it was figured out how to build an architecture for a generic data marketplace platform. This signifies the first sub-research question,

      SQ1: *How to build an architecture for a generic data marketplace platform?*

    - Then, using the methodology generating from answering *SQ1*, an architecture was built which reflected a generic data marketplace platform. This signified the second sub-research question,

      SQ2: *How does a generic data marketplace platform look like?*

- ***RT2:*** *To identify the threats associated with the architecture from RT1*. This task entails the modelling of threats around the architecture from *RT1*. This also involved subtasks which signify the same setup as *RT1*, *methodology* and *execution*. The 2 subtasks were,
    - Firstly, it was figured out how to model the threats for the architecture from *RT1* which signifies the third sub-research question,

      SQ3: *How to model the threats for the architecture of the data marketplace platform from SQ2?*

    - Secondly, using the methodology generated from answering *SQ3,* a threat model comprising of the threats associated with the architecture from *RT1* was built*;* which signifies fourth sub-research question,

      SQ4: *What are the threats associated with the data marketplace platform from SQ2?*

- ***RT3:*** *To investigate the effect of MPC technology on the architecture from RT1 and the threat model from RT2.* This involves 2 subtasks which signify the *MPC incorporation* into the architecture and effect of that incorporation on the threat model. The 2 subtasks were,
    - Firstly, it was figured out how to incorporate MPC technology into the architecture from *RT1;* which signifies the fifth sub-research question,

      SQ5: *How to incorporate MPC technology into the architecture of the data marketplace platform from SQ2?*

    - Then, the effect of MPC incorporation on the rest of the architecture and the threat model from *RT2* were deduced, which signify the sixth and seventh sub-research question as follows,

      SQ6: *What is the effect of MPC incorporation on the rest of the architecture from SQ2?*
      and
      SQ7: *What is the effect of MPC incorporation on the threats associated with the data marketplace platform from SQ4?*

The resulting artefacts from the conceptualisation phase which serve our research objective are referred as ***Artefacts 1.0*** which are listed as follows,

- *Pre-MPC Data Marketplace Platform 1.0 (SQ2)*
- *Pre-MPC Threat Model 1.0 (SQ4)*
- *Post-MPC Data Marketplace Platform 1.0 (SQ5)*
- *Post-MPC Threat Model 1.0 (SQ6)*

This was followed by the *validation phase* during which all the artefacts from the conceptualisation phase and their subsequent theoretical concepts were validated (*refined, updated, modified or invalidated*) by comparing and relating to the empirical phenomenon (*by interviewing experts; not by actually observing the phenomenon though*) to obtain more-valid artefacts serving the research objective. The validation phase was further broken down into 2 research tasks.

- ***RT4:*** *To design the methodology for conducting validation.* This entails the selection of the type of research, the research design and the formulation of the subsequent methodology as dictated by the agenda of validating the artefacts and their subsequent theoretical concepts from the conceptualisation phase. This signifies the seventh sub-research question,

  SQ8: *How to validate the artefacts and their theoretical concepts obtained from the conceptualisation phase?*

- ***RT5****: To validate the artefacts from the conceptualisation phase*. Using the methodology designed in *RT4*, the artefacts and their subsequent theoretical concepts were validated. This entails the 4 subtasks reflecting the validation and refinement of the 4 artefacts.
  - To validate the *Pre-MPC Data Marketplace Platform 1.0* and refine its theoretical concepts; which signify the following 2 sub-research questions,

    SQ9: Is the *Pre-MPC Data Marketplace Platform 1.0* valid?
    and
    SQ10: *How do the expert insights change the architecture of the data marketplace platform from SQ2?*

  - To validate the *Post-MPC Data Marketplace Platform 1.0* and refine its theoretical concepts; which signify the following 2 sub-research questions,

    SQ11: Is the *Post-MPC Data Marketplace Platform 1.0* valid?
    and
    SQ12: *What according to the experts, can be the effect of MPC technology on the architecture of the data marketplace platform from SQ10?*

  - To validate the *Pre-MPC Threat Model 1.0* and refine its theoretical concepts; which signify the following 2 sub-research questions,

    SQ13: Is the *Pre-MPC Threat Model 1.0* valid?
    and
    SQ14: *What according to the experts, are the threats associated with the data marketplaces?*

  - To validate the *Post-MPC Threat Model 1.0* and refine its theoretical concepts; which signify the following 2 sub-research questions,

SQ15: Is the *Post-MPC Threat Model 1.0* valid?
and
SQ16: *What according to the experts, can be the effect of MPC incorporation on the threats associated with the data marketplaces?*

The validated *(refined, updated or modified)* artefacts from the *validation phase* which serve our research objective are referred as ***Artefacts 2.0*** which are listed as follows,

- *Pre-MPC Data Marketplace Platform 2.0 (SQ8 & SQ9)*
- *Post-MPC Data Marketplace Platform 2.0 (SQ10 & SQ11)*
- *Pre-MPC Threat Model 2.0 (SQ12 & SQ13)*
- *Post-MPC Threat Model 2.0 (SQ14 & SQ15)*

The difference between the *Pre-MPC Data Marketplace Platform 2.0* and *Post-MPC Data Marketplace Platform 2.0* and their subsequent validated theoretical concepts signify the architectural implication of the matured MPC technology to the data marketplaces. Similarly, the difference between the *Pre-MPC Threat Model 2.0* and *Post-MPC Threat Model 2.0* and their subsequent validated theoretical concepts signify the implication of the matured MPC technology to the threat landscape of the data marketplaces. Using these differences, hypotheses are developed at the end of thesis which answers the main research question, *RQ* and fulfils the research objective, *RO*. The whole research setup discussed so far is listed in Table 1.

Table 1: *Research Setup of the Thesis*

Research Tasks	***Sub-Research Questions***	***Resulting Artefacts***
CONCEPTUALISATION PHASE		
RT1: *To build an architecture of a generic data marketplace platform*	***SQ1:*** *How to build an architecture for a generic data marketplace platform?*	***Pre-MPC Data Marketplace Platform 1.0***
	SQ2: *How does a generic data marketplace platform look like?*	
RT2: *To identify the threats associated with the architecture from RT1*	***SQ3:*** *How to model the threats for the architecture of the data marketplace platform from SQ2?*	***Pre-MPC Threat Model 1.0***
	SQ4: *What are the threats associated with the data marketplace platform from SQ2?*	
RT3: *To investigate the effect of MPC technology on the architecture from RT1 and the threat model from RT2*	***SQ5:*** *How to incorporate MPC technology into the architecture of the data marketplace platform from SQ2?*	***Post-MPC Data Marketplace Platform 1.0***
	SQ6: *What is the effect of MPC incorporation on the rest of the architecture from SQ2?*	
	SQ7: *What is the effect of MPC incorporation on the threats associated with the data marketplace platform from SQ4?*	***Post-MPC Threat Model 1.0***
VALIDATION PHASE		
RT4: *To design the methodology for conducting validation*	***SQ8:*** *How to validate the artefacts and their theoretical concepts obtained from the conceptualisation phase?*	***Validation Methodology***

***RT5**: To validate the artefacts from the conceptualisation phase*	***SQ9:*** Is the *Pre-MPC Data Marketplace Platform 1.0* valid?	***Pre-MPC Data Marketplace Platform 2.0***
	SQ10: *How do the expert insights change the architecture of the data marketplace platform from SQ2?*	
	SQ11: Is the *Post-MPC Data Marketplace Platform 1.0* valid?	***Post-MPC Data Marketplace Platform 2.0***
	SQ12: *What according to the experts, can be the effect of MPC technology on the architecture of the data marketplace platform from SQ10?*	
	SQ13: Is the *Pre-MPC Threat Model 1.0* valid?	***Pre-MPC Threat Model 2.0***
	SQ14: *What according to the experts, are the threats associated with the data marketplaces?*	
	SQ15: Is the *Post-MPC Threat Model 1.0* valid?	***Post-MPC Threat Model 2.0***
	SQ16: *What according to the experts, can be the effect of MPC incorporation on the threats associated with the data marketplaces?*	

1.3 Research Design

As already established that there is absolutely no research effort invested so far towards solving the focal research problem, the research objective of this thesis checks off all the boxes for the research to be an *exploratory study* (Sekaran & Bougie, 2013). Consequently, the research of this thesis reflects only the *theoretical framework* stage of a hypothetico-deductive research.

A theoretical framework is a representation of an observable phenomenon and an explanation on why that phenomenon is so (Sekaran & Bougie, 2013). Theoretical framework entails development of a conceptual model which embodies the theory behind the phenomenon; both of which are developed on the line of *inductive reasoning* by integrating the logical beliefs with the existing research. Following this conceptual model, hypotheses (or propositions) are developed which are the statements that explain the relationships between the different aspects of the developed theory. The hypotheses development is followed by *deduction* through which the hypotheses are tested with the empirical phenomenon by developing a way to measure the variables in the conceptual model (metrics). However, this is out of our scope as the phenomenon associated with our research objective is a future prediction and can be tested with deductive reasoning when the maturation of MPC technology happens and data marketplaces are up and running so that the implication can be measured by practically observing it. For now, the thesis delivers a theoretical framework representing the probable implications of the maturation of MPC technology to the architecture and threat landscape of the data marketplace platforms as illustrated in Figure 1.

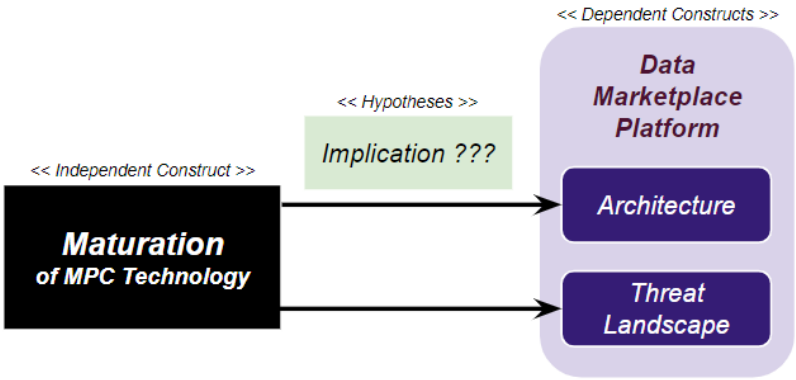


***Figure 1**: Prospective Causal Diagram resulting from the Thesis*

The updated artefacts obtained at the end of the thesis; namely, *Pre-MPC Data Marketplace Platform 2.0, Post-MPC Data Marketplace Platform 2.0, Pre-MPC Threat Model 2.0 and Pre-MPC Threat Model 2.0* constitute the conceptual models which are used towards building the theoretical framework and the subsequent hypotheses; which can fuel the next step in the empirical cycle i.e. *deduction,* associated with our focal research problem.

1.3.1 Research Framework

It is already established that the research entailed a very significant conceptualisation phase providing the first iteration of the 4 4 conceptual models developed through desk research methods. This was followed by an equally significant validation phase which involved the *validation* of the 4 conceptual models to obtain second iterations of the same. This was executed by designing the research as illustrated by the research framework in the Figure 2 which reflects all the research tasks and resulting artefacts established so far. The research methodology used for either of the phases are described in the following sub-sections.

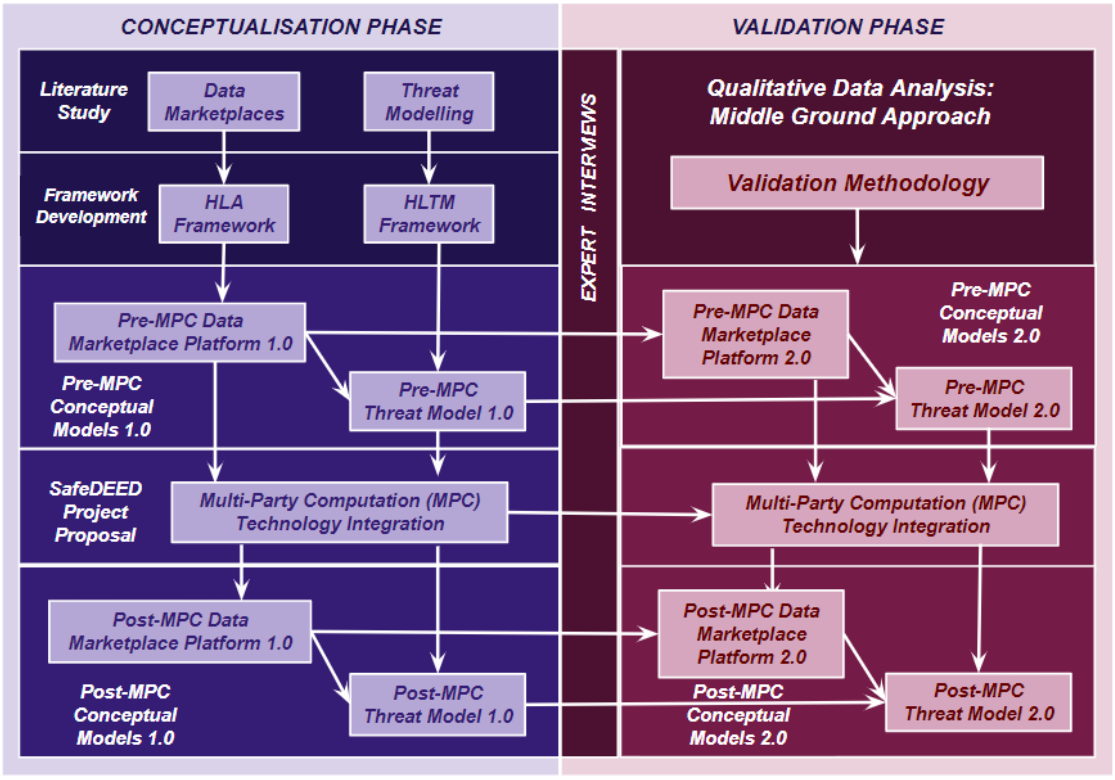


***Figure 2**: Research Framework of the Thesis*

1.3.2 Conceptualisation Design

The conceptualisation of the first iteration of the 4 conceptual models were derived solely from the desk research methods. This was done to complement the existing knowledge with our logical beliefs and experience such that all the knowledge bases at our disposal are exhausted and the further research of gaining expert insights resulted only in new knowledge. This phase was executed with the help of the following desk research methods:

- ***Literature Study:*** This method contributed significantly in laying the foundation of the thesis by supplying the existing knowledge to build all our concepts on it. The process of the literature study involved 3 steps: *searching the literature, reviewing the selected literature and critically analysing the obtained knowledge to appropriately use in building our conceptualisations*. Related to our research, literature study was conducted extensively on the subjects; data marketplaces and threat modelling while a limited amount on the cyber threats associated with the information systems and MPC technology. The methodology used to conduct the literature search is explained in detail in the subsequent chapters wherever applicable. The literature reviewed in this thesis were adopted from the sources ranging from academic (*j1ournal articles, conference articles, theses*) to non-academic ones (*consultancy articles, white papers, web articles et cetera*).
- ***Framework Development:*** This is desk research method signifying the process used to develop *HLA framework* and *HLTM framework*. These were developed respectively during *RT1* and *RT2* by understanding the fundamental concepts associated with the technological entities in general and the cyber threat modelling of those technological entities respectively.
- ***HLA Framework:*** This was used to build the high-level architecture for a generic data marketplace platform which signified the *Pre-MPC Data Marketplace Platform 1.0.* The framework is described in detail in Chapter 2.
- ***HLTM Framework:*** Similarly, this was used to build the high-level cyber threat model for the data marketplace platform consisting of the cyber threats associated with the business functions of the data marketplace platform. The framework is described in Chapter 3.

Every research activity carried out as part of the conceptualisation phase are listed in Table 2 along with their resulting artefacts.

***Table 2**: Research Activities devised for Conceptualisation Phase*

CONCEPTUALISATION PHASE			
RTs	***SQs***	***Research Activities***	***Resulting Artefacts***
RT1	***SQ1***	✓ Literature Study on Data Marketplaces ✓ *HLA* Framework Development	*High-Level Architecture (HLA) Framework;*
	SQ2	✓ Application of HLA Framework ✓ Formalisation of the Functional Requirements, Customers and Functional Components of a Data Marketplace Platform ✓ Diagrammatic Illustration of the High-Level Architecture	*A new High-Level Architecture of a generic Data Marketplace Platform* ***(Pre-MPC Data Marketplace Platform 1.0)***

RT2	***SQ3***	✓ Literature Study on Threat Modelling ✓ Literature Review of Threat Modelling Frameworks/Methodologies and Existing Threat Models ✓ HLTM Framework Development ✓ Conceptualisation of Threat Landscape	*High-Level Threat Modelling (HLTM) Framework*
			Threat Landscape Conceptualisation 1.0
	SQ4	✓ Application of HLTM Framework ✓ Literature Analysis of Cyberattack Vectors and their consequences ✓ Threat Model Development	*A new Threat Model with the high-level cyber threats to the components of the High-Level Architecture from SQ2* ***(Pre-MPC Threat Model 1.0)***
RT3	***SQ5***	✓ Study of *SafeDEED* Proposal ✓ Literature Search for the MPC Processes ✓ Formulation of MPC Incorporation	*MPC Incorporated High-Level Architecture of the Data Marketplace Platform* ***(Post-MPC Data Marketplace Platform 1.0)***
	SQ6	✓ Investigation of the effect of MPC incorporation on the architecture from *RT2* ✓ Updation of the High-Level Architecture of the Data Marketplace Platform from *SQ2* with MPC incorporation	
	SQ7	✓ Investigation of the effect of MPC incorporation on the Threat Model from *SQ4* ✓ Updation of the Threat Model from *SQ4* to reflect the effect of MPC incorporation	*Updated Threat Model reflecting the effect of the MPC incorporation* ***(Post-MPC Threat Model 1.0)***

1.3.3 Validation Design

Since the purpose of this thesis was to build a theory, it was solely concentrated on maximising the internal validity of the resulting theory and the subsequent conceptual models. Hence, the purpose of the empirical research was formalised to be the validation of the first iteration of the conceptual models resulting out of the conceptualisation phase to obtain more valid models. Additionally, the theoretical concepts associated with all the 4 artefacts was dealt and updated. Evidently the validation activity was carried out with the underlying principle which was to conduct a rigorous research to obtain precise concepts to be represented in a parsimonious way; thus, reflecting the agenda to obtain internally valid conceptual models. Furthermore, the validation activity involves refining, updated, modifying or invalidating the theoretical concepts.

The design of the validation activity which was utilised to obtain the second iterations of the 4 conceptual models is teased here. It was decided to do a *qualitative study* for this agenda as qualitative data provides the flexibility needed to carry out exploration with inductive reasoning. The qualitative data was collected with the method of *interviews* given that it provides rich primary data about the phenomenon. The prospective actor for the interviews were deduced to be the *subject area experts* as they are the only ones who possess the knowledge associated with such a cutting-edge non-mainstream research problem. As a result, *judgement sampling* was carried out to scout for the eligible experts for the study (Sekaran & Bougie, 2013). The recruited experts comprise of *researchers* and *industry experts* in the subject areas of Data Marketplaces, Threat Modelling and MPC technology. The interviews were conducted in the month of ***July 2019*** and was conducted at one-shot, i.e. the study qualifies to be cross sectional study (Sekaran & Bougie, 2013)

The research strategy for the data analysis was chosen to be *Middle-Ground Approach* (Sekaran & Bougie, 2013; de Reuver, 2019), a variant of *Grounded Theory* (a very common method used to generate theoretical frameworks (Corbin & Strauss, 1990)). The rest of the details on the validation

methodology is described in detail in Chapter 6. However, every research activity used to execute the validation phase is listed here in Table 3 along with their respective resulting artefacts.

***Table 3**: Research Activities devised for Validation Phase*

<table>
<tr><th colspan="4">VALIDATION PHASE</th></tr>
<tr><th>RTs</th><th>SQs</th><th>Research Activities</th><th>Resulting Artefacts</th></tr>
<tr><td rowspan="4">RT4</td><td rowspan="4">SQ8</td><td rowspan="4">✓ Design of the Validation Methodology
✓ Derivation of the initial set of categories and codes from the theoretical concepts developed in the conceptualisation phase.
✓ Drafting Interview Protocols
✓ Interview Scheduling
✓ Conduction of Semi-Structured Skype Interviews
✓ Recording the Interviews
✓ Transcription of Interviews Recordings
✓ Middle-Ground Approach execution
✓ Final Conceptual Model and Hypotheses Development</td><td>Interview Transcripts</td></tr>
<tr><td>Updated List of Categories and Codes</td></tr>
<tr><td>Validated Interpretations of the Theoretical Concepts associated with each Conceptual Model</td></tr>
<tr><td>Hypotheses reflecting the implication of the Maturation of MPC technology to the Architecture and the Threat Landscape of the Data Marketplaces</td></tr>
<tr><td rowspan="15">RT6</td><td>SQ9</td><td rowspan="3">✓ Validation of the Concepts of Data Marketplace Platforms
✓ Validation of Pre-MPC Data Marketplace Platform 1.0
✓ Updation of HLA Framework</td><td>Data Marketplace Platform Design Taxonomy 2.0</td></tr>
<tr><td rowspan="2">SQ10</td><td>Updated High-Level Architecture of a generic Data Marketplace Platform (Pre-MPC Data Marketplace Platform 2.0)</td></tr>
<tr><td>HLA Framework 2.0</td></tr>
<tr><td>SQ11</td><td rowspan="2">✓ Validation of the Perception of MPC Technology
✓ Validation the MPC processes suggested by SafeDEED
✓ MPC incorporation into the Pre-MPC Data Marketplace Platform 2.0.</td><td rowspan="2">MPC incorporated High-Level Architecture of the Data Marketplace Platform (Post-MPC Data Marketplace Platform 2.0)</td></tr>
<tr><td>SQ12</td></tr>
<tr><td rowspan="3">SQ13</td><td rowspan="3">✓ Validation of HLTM Framework
✓ Validation of the Conceptualisation of Threat Landscape</td><td>Threat Model Taxonomy 2.0</td></tr>
<tr><td>Threat Landscape Conceptualisation 2.0</td></tr>
<tr><td>HLCTM Framework</td></tr>
<tr><td>SQ14</td><td>✓ Generation of new business threat model for the Data Marketplaces using the insights generated from the expert interviews.</td><td>A New Threat Model with the high-level business threats associated with the Data Marketplaces (Pre-MPC Threat Model 2.0)</td></tr>
<tr><td>SQ15</td><td rowspan="2">✓ Analysis of the effect of MPC by examining Post-MPC Data Marketplace Platform 2.0 and its impact Pre-MPC Threat Model 2.0</td><td rowspan="2">The Threat Model consisting of the business threats which apply even after the incorporation of MPC technology into the data marketplace platform (Post-MPC Threat Model 2.0)</td></tr>
<tr><td>SQ16</td></tr>
</table>

Though, the underlying principle of the validation methodology was to conduct rigorous research to obtain precise results and further represent them in a parsimonious way, *theoretical saturation* was not achieved for our Middle-Ground Approach, owing to the constraints of the time and

unavailability of enough sample size of the experts., As a result, the findings of this thesis represent the initial iterations of the empirical cycle associated with the research problem. On the brighter side, this thesis represents the pioneer effort towards addressing the focal research problem. Hence, even though, the findings are relatively less mature, we leave behind a myriad of research opportunities related to the focal research problem, be it directly continuing the research or going beyond with a different focus.

1.4 Scope of the Research

The underlying specification of the research is that everything in the research are analysed from the perspective of Management of Technology. So, the low-level technical specification of any aspect is not considered, and the research is carried out from a technology manager's perspective.

The generic nature of the research problem and the research objective makes it possible to carry out the research at various scopes, either at the technological level or business level or enterprise level or even data market level. However, because of the time constraint and the availability of the relevant resources, a limited scope of analysing data marketplaces at a technological level was formulated such that credible results could be obtained within the stipulated time. However, the scope changed during the course of the research from technological level to the business level because of the reason discussed in the rest of the section

Initially in the *conceptualisation phase*, the scope with respect to research objective of the thesis was considered to be at the technological level of the data marketplace platforms. This means that the data marketplaces were analysed from the perspective of them being technological platforms and hence, the architecture was considered to be of a technological system consisting of individual components made up of information systems. This made sense for the MPC technology also, as it is a technology which is incorporated technologically, thus reinforcing our focus of analysing at the limited scope of technological level.

However, this interpretation was rejected during the expert interviews during *validation phase*. It was remarked by the experts that data marketplaces cannot be materialised just by technology but should also involve the non-technological element consisting of governance model which comprises of legal entities, auditing authorities and other crucial human actors. It was deduced that even these non-technological elements are potentially influenced by the incorporation of the MPC technology. Hence, the scope was changed from seeing data marketplaces as technological platforms to seeing them as businesses whose realisation is influenced by the right combination of the technological and non-technological elements. This helped in gaining a comprehensive perspective with respect the implication of the MPC technology towards the data marketplaces as business entities.

This change of scope also happened with respect to threat landscape during the *validation phase* but for a different reason. To investigate the threat landscape, initially, it was conceptualised to focus on the cyber threats to information systems within each component and to sum up all the business consequences to collectively represent the threat landscape. But during the expert interviews, it was remarked by the experts that it is crucial to analyse the threat landscape at a business level than the technology level to understand the true implication of the MPC technology to the data marketplaces. The reasoning was that the cyber threats could be mitigated by security technologies; but the potential of MPC technology does not just solve the threats at the technological level but especially at the business level which are complex to solve and pose threat to the business logic of the data marketplaces. As a result, in order to obtain credible results contributing to our agenda, the scope was widened where the data marketplaces were analysed at the business level and threats to the business logic of the data marketplaces were deduced to

establish the threat landscape. This provided relevant results which helped in explicating the implication of MPC technology to the actual threat landscape suffered by data marketplaces.

The widening of the scope is reflected in our 2 iterations of the conceptual models; where the first iteration comprising of ***Artefacts 1.0*** represent the conceptual models at the scope of technological level. On the other hand, the second iteration comprising of ***Artefacts 2.0*** represent the conceptual models at the scope of business level. The scope described so far reflects the high-level scope of the research. The detailed specification with respect to different aspects of the research is described in the comping subsections.

1.4.1 Scope for the Data Marketplaces

It is already mentioned that data marketplaces are initially analysed as technological platforms and then the focus changes it to be business species. Here, we discuss about the further aspects crucial to describe. Any specific data marketplace was not considered as a basis for the analysis in this research. Instead an abstraction of a generic data marketplaces is created, and further analysis is performed with respect to the same abstraction.

Firstly, the data is viewed here as a tradable commodity and the focus of the research remains the same throughout. As a result, the definition of data, the different types of data, the content present in the data et cetera are out of our scope. The data considered here is just commercial data which may or may not contain personal information depending on how the data was collected by the company who wishes to commoditise that data. Though the data and its types are part of a significant research area, that aspect is not part of the scope of our research.

With respect to the type of data marketplace designs, the focus improved as and when relevant information was uncovered. Firstly, the type of data marketplaces was established to be *Business (B2B) data marketplaces* as our thesis is aiming to contribute towards SafeDEED's agenda of inter-organizational data sharing. Then, with respect to the platform design, it was focussed to be just *many-to-many* or *multilateral B2B data marketplaces* from the classification of Koutroumpis et al. (2017). Then, after the findings from the interviews and further analysis in *Validation Phase*, we expanded the classification and refined our design to be *Many-to-Many B2B Decentralised Serendipity Model data marketplaces* as it represented the most generic form of data marketplaces that practically exist.

Coming to the architecture of the data marketplace, initially we wished to analyse from the surface level of the technological level which gave rise to the conceptualisation of High-Level Architecture. Later, the high-level architecture concept also underwent change and widened its scope to incorporate the non-technological aspects as mentioned earlier. To conceptualise the actors involved in the data marketplaces, only customers were included during the conceptualisation phase but were expanded to involve the ecosystem of the data marketplaces to have a comprehensive overview of the actors at the business level. The first iteration of these aspects is dealt in Chapter 2 and then, widened in Chapter 7 as part of the HLA framework formulation.

1.4.2 Scope for the Threat Modelling

The same issue of expanding of the scope happened here too. Initially in the conceptualisation phase, the threats were focussed to be cyber threats acting at the technological level which were described at a high-level without the detailed description of the threat scenario. These kinds of cyber threats were referred as High-Level Threats. But the focus was later changed to business threats which act to the data marketplaces at the level of its business logic. This expansion of the scope and its further specification can be found in the descriptions of HLTM framework in Chapter

2 and HLCTM framework Chapter 7. Furthermore, the threats at the business level can comprise of business threats and legal threats. Since legal domain is not our expertise, we limited our scope only to the *business threats.*

1.5 Knowledge Contribution

This thesis mainly provides a theoretical framework with conceptual models and corresponding hypotheses reflecting the research objective towards solving the focal research problem. Parallelly, the thesis also aims to fill the knowledge gaps associated with the subject areas: *Data Marketplaces, Threat Modelling* and *MPC Technology*.

- ***Data Marketplaces***: The phenomenon of the data marketplaces is a recent development which is gaining momentum. As a result, there are a lot of gaps associated with the literature on the data marketplaces; some of which are associated with the concepts like their architecture design, business models et cetera. With Europe currently undergoing a transition towards data-driven economy and because of the benefits offered by the data marketplaces for the data-driven economy, the research on data marketplace has become significantly relevant for the researchers to explore this area. For the same reason, we also carried out this research so that we could fill the gap to the extent of our best abilities. The contribution ranges from the technological architecture to business architecture of the data marketplaces which do not exist in the literature currently.
- ***Threat Modelling:*** This area is fairly familiar, and a lot of research exists already. However, threat modelling of our focal entity, data marketplaces has never been carried. Most of the existing threat modelling literature is directed towards the area of software engineering or at the level of information systems (which is also referred to as cyber threat modelling). There not much research with respect threat modelling at a high-level and this thesis contributes to this agenda by providing a higholevel threat modelling framework which contributes towards filling the gap of threat modelling at the level of business functions.
- ***MPC Technology***: The work on this concept has been going on for a long time in the research community. However, the technology has not yet matured (scalability issue) enough to find real-life applications. SafeDEED has taken upon itself to find solutions to the limitations of MPC technology as its core task and thereby, propose to make the technology mature for real-life application. We are building on the above this proposed claims to find potential application for MPC technology in the unexplored species of the data marketplaces. By this, we aim to fill the gap associated with the business application of MPC Technology.

1.6 Structure of the Thesis

The thesis is structured into 4 parts: *Introduction, Conceptualisation, Validation* and *Conclusion.* The constituents of these parts are listed in Table 4.

Table 4: Structure of the Thesis Report

<table>
<tr><td colspan="3">INTRODUCTION</td><td colspan="2">Chapter 1: Introduction
Research Problem</td></tr>
<tr><td rowspan="8">CONCEPTUALISATION</td><td rowspan="2">RT1</td><td>SQ1</td><td colspan="2">Chapter 2: A Study on Data Marketplaces</td></tr>
<tr><td>SQ2</td><td>✓ Challenges of Commoditising Data
✓ Data Marketplace Platform Designs
✓ Application of HLA Framework
✓ Functional Requirements
✓ Customers
✓ Functional Components</td><td>HLA Framework
Pre-MPC Data Marketplace Platform 1.0</td></tr>
<tr><td rowspan="2">RT2</td><td>SQ3</td><td colspan="2">Chapter 3: A Study on Threat Modelling</td></tr>
<tr><td></td><td>✓ Process of Threat Modelling</td><td>Threat Landscape Conceptualisation 1.0
HLTM Framework</td></tr>
<tr><td rowspan="3">RT3</td><td>SQ4</td><td colspan="2">Chapter 4: A New Threat Model for Data Marketplace Platforms</td></tr>
<tr><td>✓ Application of HLTM Framework</td><td>Pre-MPC Threat Model 1.0</td></tr>
<tr><td>SQ5</td><td colspan="2">Chapter 5: Effect of MPC on Architecture and Threat Landscape of Data Marketplaces</td></tr>
<tr><td>SQ6</td><td rowspan="2">✓ Concept of Multi-Party Computation (MPC) Technology
✓ MPC Incorporation into Pre-Data Marketplace Platform 1.0
✓ Effect of MPC Incorporation on Pre-MPC Threat Model 1.0</td><td>Post-MPC Data Marketplace Platform 1.0</td></tr>
<tr><td>SQ7</td><td>Post-MPC Threat Model 1.0</td></tr>
<tr><td rowspan="10">VALIDATIONH</td><td rowspan="2">RT4</td><td>SQ8</td><td colspan="2">Chapter 6: Validation Methodology</td></tr>
<tr><td>✓ Design of the Middle-Ground Approach for Qualitative Data Analysis of Expert Interviews</td><td>Validation Methodology</td></tr>
<tr><td rowspan="8">RT5</td><td>SQ9</td><td colspan="2">Chapter 7: Results and Analyses</td></tr>
<tr><td>SQ10</td><td rowspan="7">✓ Execution of Validation Methodology
✓ Validation and Updation of Conceptual Models 1.0 and their corresponding theoretical concepts from the Conceptualisation Phase</td><td>HLA Framework 2.0</td></tr>
<tr><td>SQ11</td><td rowspan="2">Pre-MPC Data Marketplace Platform 2.0</td></tr>
<tr><td>SQ12</td></tr>
<tr><td>SQ13</td><td>Post-MPC Data Marketplace Platform 2.0</td></tr>
<tr><td>SQ14</td><td>Threat Landscape Conceptualisation 2.0</td></tr>
<tr><td>SQ15</td><td>HLCTM Framework</td></tr>
<tr><td>SQ16</td><td>Pre-MPC Threat Model 2.0</td></tr>
<tr><td>CONCLUSION</td><td>RO</td><td>RQ</td><td colspan="2">Chapter 8: Conclusions and Discussions</td></tr>
<tr><td colspan="3">✓ Theoretical Framework and Hypotheses Development
✓ Answers to RQ and SQs
✓ Contributions
✓ Limitations
✓ Future Work Recommendations</td><td>Implication of the Maturation of MPC Technology to the Architecture and the Threat Landscape of the Data Marketplace Platforms</td></tr>
</table>

2

A Study on Data Marketplaces

This chapter marks the start of the *Conceptualisation phase* which involved building the theoretical concepts associated with the research tasks *RT1, RT2 and RT3.* In this chapter, the research task, *RT1: To build an architecture of a generic data marketplace platform,* is dealt, and the following 2 sub-research questions are answered.

SQ1: How to build an architecture for a generic data marketplace platform?
and
SQ2: How does a generic data marketplace platform look like?

A literature study was conducted on data marketplaces with an aim to explore the phenomenon of data marketplaces and to understand their fundamental concepts like the definition, different features, relevant actors et cetera. Following this, a new framework was developed to build an architecture for a generic data marketplace platform which answers, *SQ1*. Then, the framework was applied to obtain an architecture of the data marketplace platform, which answers *SQ2*.

The rest of the chapter is structured as follows. Section 2.1 describes the methodology used to search and select the relevant literature on data marketplaces. Section 2.2 describes what makes data a unique commodity and its related challenges. Section 2.3 provides an overview of the data marketplaces comprising of its definition, types and different platform designs. Section 2.4 describes the *High-Level Architecture (HLA) framework.* Section 2.5 depicts the application of the HLA framework and the conceptualisations of *functional requirements, actors* and *functional components* to obtain the *Pre-MPC Data Marketplace Platform 1.0*. Section 2.6 summarises the chapter where the focal sub-research questions, *SQ1* and *SQ2* are formally answered.

2.1 Literature Search and Selection Methodology

The focus of this literature study was to obtain an architecture of a generic data marketplace platform. The criteria formulated for the search and selection of the literature was that the literature should comprise of the fundamental concepts associated with the data marketplaces like its characteristics, the basic architecture, the functionalities and features, the actors in its ecosystem

and further concepts on these lines; so that using these concepts, an architecture for a generic data marketplace platform can be established

To pursue this agenda, a simple search was performed on *Web of Science* with the search phrase, "***data marketplaces***" which resulted in ***19*** articles. The same search on *Scopus* yielded ***69*** articles which also consisted of all the ***19*** articles found previously on Web of Science. Hence, further search was performed only on Scopus owing to its richness. Later, a filter to exclude the articles from the conference proceedings was applied; so that only the literature of high quality was considered like the articles from peer-reviewed sources. The filter yielded ***18*** articles. After a quick scan of the title and abstract of the articles, it was found that few of the articles dealt with specific issues in the area of data marketplaces like data pricing (Muschalle et al, 2013; Fricker & Maksimov, 2017), metadata (Spiekermann et al, 2018) et cetera; while most of the articles proposed many data marketplaces for specific domains like automobile industry, health care industry, credit scoring et cetera where data can be shared among incumbent actors to obtain benefit from the data within the industry. The functionality aspect of the data marketplaces was mentioned in a very few articles and even those articles, it was not dealt with more focus. Later, the search was expanded to include *conference articles* and more sources like *white papers, consultancy literature* et cetera by performing the same search on *Google Scholar*. This gave an enormous number of results related anywhere near to data marketplaces; which also included the articles from the previous searches on Web of Science and Scopus.

Although there is a reasonable amount of literature related to data marketplaces in general, with most of them dealing with the pricing techniques of data products, there is a scarcity of literature related to the basic functioning of data marketplaces. The reason can be that initially, the focus was on setting up a data marketplace, and figuring out how to price the data. Only recently, with the events of data marketplaces failing (Schomm, Stahl, & Vossen, 2013) or stopping their operations (Ramel, 2016), the researchers could have gotten interested to investigate the issues with their functioning. Hence, the literature on the functioning of data marketplaces is still in its infancy. Also, the advent of BlockChain and other enabling technologies crucial for the functioning of the data marketplaces, happened just recently. Hence, the research on this agenda has picked up momentum only recently. Because of the scarcity of literature on the subject matter, the selection of literature was done based on the availability of relevant information rather than judging the quality of the literature. Although the relevance of the information is critically analysed throughout the study wherever applicable. This literature search was conducted till ***10 May 2019***. Any literature published after this date was not considered to be part of this study.

The search results were examined to find if they fitted the search criteria. The stages of filtration performed were based on: firstly, the title, then the abstract and then the skim-read understanding of the articles. After this filtration, 4 categories of literature were selected.

- Firstly, the works of Florian Stahl, Fabian Schomm and a few more collaborators (Muschalle, Stahl, Löser, & Vossen, 2013; Schomm et al., 2013; Stahl, Schomm, & Vossen, 2014; Stahl, Schomm, Vossen, & Vomfell, 2016; Stahl, Schomm, Vomfell, & Vossen, 2017) who have studied the phenomenon of data as a commodity, conducted surveys on real-life data marketplaces and provided a classification encompassing all kinds of data marketplaces. These were the pioneers in the research of data marketplaces and hence, were included in our study.
- Secondly, the work of Pantelis Koutroumpis (Koutroumpis & Leiponen, 2013; Koutroumpis, Leiponen, & Thomas, 2017) who has studied big data and data marketplaces from an economic perspective. His works fit the search criteria for they contain the basic concepts associated with the data marketplaces like their business logic, challenges involved in setting up a data marketplace.
- Thirdly, the research conducted by Fraunhofer Institute for Applied Information Technology (Quix, Chakrabarti, Kleff, & Pullmann, 2017; Chakrabarti, Quix, Geisler, Khromov,

& Jarke, 2018) who are conducting research on data marketplaces to set up their own data market ecosystem called Industrial Data Space. Their work contributes to our criteria by providing concepts related to the architectural aspects like the functionalities and feature, the actors et cetera.

- Finally, a set of articles which provide secondary and tertiary information about data marketplaces like big data, data commercialisation, data contracts, metadata models etc which were identified through backward and forward snowballing of the above-three categories were included in the study.

2.2 Data as a Commodity

To understand what makes the data marketplaces a unique species of business, it is important to understand the marketplace's commodity, i.e. data. Data, as a trading commodity, exhibits very different characteristics than a normal good. These characteristics pose challenges for the successful commodification of data. The challenges can be twofold which are termed as *Weak Protection Regime and Data Sharing Reluctance* which is discussed as follows.

2.2.1 Weak Protection Regime

Koutroumpis et al., (2013) suggests that data belongs to the category of goods called *non-rivalrous goods.* These goods can be replicated with negligible cost and the same good can be used simultaneously at multiple locations by different entities. Furthermore, Koutroumpis et al. (2017) suggest that data is an *intermediate good* which means that it is of less or no business value unless either subjected to analysis or when combined with data from other appropriate sources, thereby creating meaningful data products. Because of these 2 characteristics, it is difficult to assign intellectual property (IP) rights to effectively protect data. The copyright laws and the database rights protect data in the confines of a database, but neither the actual data contents nor their intangible knowledge (Koutroumpis et al., 2017). So, once the data is out of the database and is modified either subjecting to analysis or combining with other datasets, the rights do not apply to the resultant data content or the extracted knowledge; and hence, it becomes almost impossible to trace the path travelled by a data point (Koutroumpis et al., 2017) (also called as *Data Lineage* which will be explained later in Chapter 7). This condition results in a *weak protection regime* which makes data a tricky commodity for trading.

2.2.2 Data Sharing Reluctance

Koutroumpis et al. (2013) categorise data as an *experience good* where the buyer has less insight in the good than the seller and sometimes, it is difficult for the seller himself to judge the value of the data. Furthermore, data suffers from *Arrow's paradox*. This is a paradoxical phenomenon where the value of the data can be convinced to the buyer only after disclosing the data; however, after the reveal, the data loses its value because of its non-rivalrous nature (Arrow, 1972). As a result, high-value data face difficulty for being transacted. Another barrier is with respect to the uncertainty associated with the regulatory space around data. The ignorance of the sellers with the respect to the regulations for data like General Data Protection Regulation (GDPR) results in the sellers being unclear or unaware of the legal status of the data. Because of these concerns, the sellers may end

up not sharing their data or may share low-quality data. We term this phenomenon as *data sharing reluctance*.

2.2.3 Implication of these Challenges

As a result of the challenges posed by the *weak protection regime* and *data sharing reluctance*, in order to be a successfully tradable commodity, data is expected to be coupled with the information about its *provenance* which includes its origin, history and properties (Koutroumpis et al., 2017). This information can be referred to as the "*metadata*" of the data good which helps in judging the credibility, quality and security status of the data. There has been considerable research on designing metadata models for data products (Koutroumpis et al., 2017; Spiekermann et al., 2018). At present, the provenance of data ends within the boundaries of the sellers and once, it is transacted, the provenance and hence, the control over the data is lost. Essentially, the data marketplaces should have a mechanism to address these challenges for their realisation.

2.3 Overview of Data Marketplaces

The concepts acquired from the literature search on the topic of data marketplaces is dealt here starting from their definition, the issues involved in materialising them and their variants; which are discussed in following subsections.

2.3.1 Definition of Data Marketplaces

The research on data marketplaces started in the form of conducting surveys of existing data marketplaces. Schomm et al. (2013) performed the first systematic survey about the data marketplaces. They define data marketplaces as platforms where registered data providers can upload and maintain datasets; while the data consumers are granted access to access and use that data through licensing models (Schomm et al., 2013). The criteria for the inclusion of data marketplaces in their survey was that the entity should provide an infrastructure for data trading. So, even the companies who just sold their data over the internet also qualified as data marketplaces. But this contradicts their definition of data marketplaces being a platform where both the sides of the data market meet. They admit this in their later work, (Stahl et al., 2015) that data marketplace platforms constitute only a category of the ones considered in their surveys. The reason for this inconsistency is the criteria of data trading infrastructure which was considered for the inclusion of companies into the survey. As a result, even the data vendors who sell data on their e-commerce websites also qualified for the survey; which evidently contradicted with their definition of data marketplaces being just platforms.

Deichmann et al., (2016) provide a more accurate definition as part of their research at McKinsey. They define data marketplaces as "*platforms that connect providers and consumers of datasets and data streams, ensuring high quality, consistency and security. The data suppliers authorize the marketplace to license their information on their behalf following defined terms and conditions*". They define this with respect to IoT data, but the definition holds good for any form of data as the focus is on the data being a commodity but not on its different types.

Koutroumpis et al. (2017) provide an even more comprehensive overview of data marketplaces by compiling observations from different sources. They classify all kinds of data marketplaces similar

to that surveyed by Schomm et al. (2013); but do so in a conceptual way and not by observation. This means that their classification consists of theoretical frameworks of different concept data marketplace platforms but not the ones that already exist in real-life. However, there relate these conceptualisations with real-life data marketplaces. In their classification, only one category reflects the true platform version of the data marketplace where any data supplier can upload and sell data to any data consumer. They call this variant as ***many-to-many*** or ***multilateral*** data marketplaces which is the focal data marketplace platform considered for this research.

Koutroumpis et al. (2017) define *multilateral* data marketplaces as *multi-sided platforms* where a digital intermediary connects data sellers, data buyers and facilitates data sharing activities. This definition is consistent also with that of Deichmann et al., (2016). Furthermore, Koutroumpis et al. (2017) go on saying that this variant does not possess the ownership of the data goods being transacted but merely orchestrate, the data exchange process through services of search/discovery, transaction validation, transaction history and payment gateway. Functionally speaking, multilateral data marketplaces enable the association of disparate datasets from different data owners through easy search and discovery, standardization of their formats and their subsequent aggregation into meaningful data products (Koutroumpis et al., 2017). This mandates the necessity of regulatory environment, communication standards, data protocols and procedures of data import, storage, transformation, aggregation, analysis and delivery functionalities (Koutroumpis et al., 2017). With these services, like any other digital marketplace platform, data marketplaces create value for their customers in the following ways as suggested by Smith et al. (2016). Firstly, the search process for data is simplified without having to browse each data provider's offering at their websites. Secondly, access to rich data content which can be compared with each other to make an informed decision. Thirdly, automated data exchange with standardised data formats makes the trading process easier. Finally, there is a larger scope for building relationships by an improved match between supply and demand of data. Smith et al. (2016) discussed these with respect to open data marketplace platforms; but can also be applicable to commercial data marketplace platforms

Despite these advantages, there exist only a few examples of functional data marketplace platforms. Recently, Microsoft Azure Data Marketplace, which was the first mover to establish data marketplace platform, closed its operations and transformed itself into a different marketplace providing sophisticated data products and analytics services; instead of just data. Microsoft mentioned that the reason for this was the lack of a customer base interested in using Microsoft Azure Data Marketplace as mentioned by Ramel (2016) in his web article. This can be attributed to network effects experienced by data marketplaces which mean the value of data marketplaces decrease if the number of participants decreases (referred to as *positive externalities*) (Eisenmann, Parker, & Van Alstyne, 2006). The reason for the customers not opting for participating can be that the existing data marketplace platform does not effectively address the challenges of commoditising data as mentioned in section 2.1; resulting in a lesser trust to trade high-value commercial data. Hence, we can find many open data marketplaces in existence which offer data of lower value (*open data*); while a very small number of commercial data marketplaces.

2.3.2 Types of Data Marketplaces

Based on the type of data and the parties involved in the exchange of data, Smith (2018), a founder of data marketplace called DX network, classifies data marketplaces in one of his web articles into 3 categories; namely,

- ***Personal Data Marketplaces***: These enable individual consumers to monetize their data by providing a platform for them to sell their data on their own terms to the concerned buyers. The individuals are provided with mobile application which collects data like social media

streams, location etc with an interface to manage the data trading activities. Some of the examples are Datum, DataWallet, Fysical etc.

- ***Business Data Marketplaces***: These enable business-to-business (B2B) data exchange by providing a platform for companies to trade data. These data marketplaces help in overcoming the differences in the formats of data handled by sellers and buyers by providing a common data model and interface to trade data.
- ***Sensor Data Marketplaces****:* These provide real-time data streams from remote sensor devices (IoT) which are listed by sellers on the platform. The type of data includes weather data, pollution data, manufacturing equipment data etc. Some of the examples include IOTA DataMarket, DataBroker DAO, Steamr etc.

The research in this thesis is focussed only towards *Business (B2B) Data Marketplaces* as the context of our research is inter-organizational data sharing specified by SafeDEED. So, the focal data marketplace platform is refined to be a "***many-to-many B2B data marketplace***" which is what we decided mean when we refer to the term "*data marketplace(s)" or "data marketplace platform(s)"* in the rest of the research. However, the focal data marketplace platform undergoes one more change during the *empirical research phase.*

2.3.3 Data Marketplace Platform Designs

Koutroumpis et al. (2017) propose 3 requirements a data marketplace should possess to overcome the previously-discussed challenges of commoditising data. They refer to these requirements as *institutional requirements* which are listed as follows,

- Strict ***boundary conditions*** to data marketplace platforms are instrumental in allowing only legitimate users to participate in the data transaction while filtering out unreliable users.
- ***Rules of usage*** enable control over data for the data sellers through data contracts which specify the criteria for data usage, thus providing legal cover restricting the misuse of data.
- ***Monitoring mechanism*** oversees all the data transactions and operations on the data marketplace platform and can detect any anomalous activity. This basically constitutes the governance aspect of the data marketplace platforms.

Based on these concepts, Koutroumpis et al., (2017) suggest 3 designs of *multilateral* data marketplace platform and discuss their relevance with respect to the above-mentioned institutional requirements. The variants are discussed as follows,

- ***Centralised platform*** hold data centrally and offers its services on a central technological platform. Koutroumpis et al. (2017) conceptualised that these platforms enforce strong *boundary conditions* through formal entrance policies but fail with respect to *rules of usage* and *monitoring mechanism*. The latter 2 are ineffective for the reason that once, the data leaves the platform, the provenance and the control over data are lost. As a result, there is no way to track or monitor the usage of data by concerned consumers. Hence, they are suitable for trading low value data like open data and hence, the open data platforms that we generally come across are *centralised platforms*.
- ***Decentralised platform*** enabled by Distributed Ledger Technology (DLT) where the data is held in a blockchain. Koutroumpis et al. (2017) suggest from their conceptualisations that this platform design enforces all the institutional requirements as follows. It diffuses the need for *boundary conditions* because of the transparent philosophy of DLT. It addresses the *rules of usage as* every transaction and usage are recorded on the ledger enabling the data owners to track the *usage* of their data points where they can detect the unusual

activities. The *monitoring mechanism* is enforced technologically by DLT. Hence, the decentralised platform design provides an effective data marketplace platform enabling the trading of high-quality data. But this design suffers from technological immaturity as the DLT is not scalable for large scale operations (Simonite, 2016). There is a considerable amount of research going on to make this design a reality. Some of the concept platform designs are Enigma (Zyskind, Nathan, & Pentland, 2015), Sterling (Hynes, Dao, Yan, Cheng, & Song, 2018), Trusted Data Marketplace (Roman & Stefano, 2016) etc.

- ***Collective platform*** is a platform design which achieves the enforcement of institutional properties by forming a closed consortium of partners (*boundary conditions*) to exchange data among each other which will be powered by complex contracts (*rules of usage*) and effective *monitoring mechanism* taken care of by a separate dedicated actor, *platform provider* (Koutroumpis et al., 2017). However, this design suffers from economic consequences like high transaction costs making them ineffective for large-scale multilateral data trading. Collective platforms are effective when they are formed by small number of partners with pre-existing trust-based relationships and shared interests of data exchange (Koutroumpis et al., 2017).

Although promising, these 3 variants are just predictive conceptualisations which practically are not fully functional. Also, the focus of Koutroumpis et al. (2017) to conceptualise these platform designs is only with respect to the institutional requirements: *boundary conditions, rules and monitoring*. But these requirements are not exhaustive as they consider only economic perspective. Apart from these, there are also additional requirements which specify further necessary aspects of data marketplace platforms. These will be dealt comprehensively in subsection 2.4.1. Furthermore, there has been advent of cutting-edge technologies like BlockChain, Multi-Party Computation (MPC), Homomorphic Encryption et cetera which can overcome the above-discussed constraints technologically alone. But this is just a claim as the said-technologies have not achieved the desired level of sophistication to be applied in real-life cases. Evidently, investigating this claim is part of our research problem but we are only doing it for MPC technology.

2.3.4 Reflection on the Literature Study of Data Marketplaces

The existing literature on the data marketplaces deals extensively with materialising the idea of the data marketplaces; which basically involves its value proposition, potential and challenges. Consiquently, there is a huge gap in the literature with respect to the data marketplaces from a functional standpoint. Although this perspective has been touched upon in few articles, there is no comprehensive understanding of the fundamental functionalities and features of a data marketplace platform, let alone an architecture for the same. Hence, we decided to build our own architecture which could represent a generic data marketplace platform. However, building an architecture for a data marketplace platform is a research problem in itself as data marketplace platforms can be implemented with a myriad of technical specifications and building such kind into a single technical architecture reflecting all the aspects comprehensively deserves a separate thesis with considerable research effort and time. However, we figured out a solution for this issue. Instead of building a full-fledged technical architecture with detailed low-level specification, it was decided to build a ***high-level architecture*** for a generic data marketplace platform reflecting the surface-level (high-level) information of data marketplaces. Though it doesn't contain technical (*low-level*) specifications to represent an accurate data marketplace platform, the high-level architecture can still be a representative of the credible phenomenon associated with the data marketplaces which could provide a comprehensive understanding on their functional aspects. This decision made sense not only because we found a few relevant articles which helped us to do build the high-level architecture (like institutional requirements for data marketplace platforms by Koutroumpis et al. (2017); goals of data marketplaces by Chakrabarti et al. (2018); Enterprise Data

Marketplace (EDM) conceptualisation by Wells (2017)); but also helped us in establishing a reduced scope for the subject of data marketplaces in our already monumental research problem. The resulting high-level architecture was expected to fill the knoweldge gap existing in the literature related to the functional and architectural aspects of the data marketplace platforms.

2.4 High-Level Architecture (*HLA*) Framework

Following the decision of building the high-level architecture of a data marketplace platform, a simple framework was formulated which could help in building the same. Since our focal entity is the species of data marketplace platforms which is a technological entity, we decided the scope of the potentially resulting archietcture to be technological which means that the resulting architecture would be a technological architecture of the data marketplace platforms but only representing their surface-level (high-level) infomation with no technical (low-level) specification. Following this scope formulation, the consitituents of the framework were formulated. For any technology, the underlying principle is that the customers of the technology dictate what the technology should deliver. Hence, as a norm for developing any technology, firstly, the requirements of that technology are specified; then the profiles of the consumers who potentially use the technology are designed and finally, the surface-level (high-level) components are decided which reflect the fundamental functionalities of the focal technology satisfying its previously-specified requirements. This philosophy holds good not only for a technology but also to a wider scope till the level of organizations. Hence, this framework is not just specific to data marketplace platform but also other business entities ranging from a simple information systems to complex organizations.

Based on the above motivation, the attributes of the framework were formalised which are listed as follows,

- ***Functional Requirements*** specify the basic requirements which are required to ensure the basic functioning of the the focal entity. These requirements can be specified at the surface level without going into any detail to support the high-level philosophy of the framework.
- ***Customers*** signify the customers who use the offering of the focal business entity. Although a non-technological attribute, the customers form a crucial ingredient as they are the ones using the technology for their benefit. Hence, it is necessary to define the customer profiles who utilise the the technology.
- ***Functional Components*** represent the block box versions of all the components of the focal entity which embody the fundamental functionalties and features which satisfy the previously-developed functional requirements Consequently, when building the high-level architecture, the functionalities and the features have to be formulated.

The *HLA framework,* as illustarted in Figure 3, reflects the answer for the sub-research question, *SQ1* as it can be used to build a high-level architetcure for the data marketplace platform. Essentially, *HLA* framework qualifies as a desk research method used to execute the second-half of *RT1.*

***Figure 3**: High-Level Architecture (HLA) Framework*

2.5 High-Level Architecture of a Data Marketplace Platform

The *HLA framework* was applied to build a high-level architecture of a generic data marketplace platform and the same is discussed in this section, As specified earlier, the resulting architecture will be a technological archietcture of the data marketplace platform with surface-level (high-level) infromation with no technical (low-level) specification. This resulting architecture answers the sub-research question, *SQ2*. The application of the *HLA framework* involves populating the values for the attributes: *functional requirements, actors* and *functional components* with the information either from the literature or further conceptualisations as applicable for our focal data marketplace platform (*many-to-many B2B data marketplace*).

2.5.1 Functional Requirements of the Data Marketplace Platform

The functional requirements here were encountered in the literature analysis in the form of,

- the *institutional requirements* of a data marketplace platform suggested by Koutroumpis et al. (2017);
- the *goals* of the data marketplace developed by Chakrabarti et al. (2018) for their Industrial Data Space project.

These requirements were analysed, and the appropriate ones were either adopted directly or interpreted as applicable to the focal data marketplace platforms of this research. These constitute necessary conditions for the basic functioning of a data marketplace and are listed and described as follows:

- ***Boundary Conditions***: Strict boundary conditions help in authorising only the legitimate participants willing to share or buy data. This helps in safeguarding the data from unauthorised access from malicious sources.
- ***Data Provenance***: The lineage of data should be tracked and the change of ownership of each data point in the offering should be documented. The provenance information is the "metadata" of the data product and the platform should have a feature to manage this metadata which helps in preserving the legal usage of data.
- ***Data Governance***: This requirement is a way of governing the trading of data by having mechanisms for management and maintenance of data, traceability of data exchange and data use.

- ***Data Economy***: This requirement simply reflects the business purpose of the data marketplace platform which is to generate revenue stream for itself through its services. Usually, this is achieved through the commissions earned from the data marketplace platform services or by further additional means.
- ***Data Sovereignty***: The platform should have mechanism for the data provider to have control over his dataset, which can be enabled by handling permissions, usage restrictions, data contracts etc or through technological solutions like Blockchain. By this, the provider can protect the legality of the data and not be worried about it being misused by the data consumer.
- ***Secure Data Exchange***: This is a requirement which relates to the most fundamental aspect of the data marketplace platform, the data exchange. The data exchange should happen in the most secure way because the data being exchanged is of high commercial value. The disclosure of such data will reduce its value and result in commercial, reputational and regulatory losses to the data actors. Hence, data exchange from the origin of data (data provider) to the actual point of use (data consumer) should happen in a secure way.
- ***Data Exchange Platform***: This is a complementary requirement resulting from combining all the previous requirements which is to have a fully equipped data exchange platform which could enable the data actors to trade data.

Among these requirements, the *boundary conditions, data provenance, data sovereignty* and *data governance* collectively were inspired from the institutional requirements as suggested by Koutroumpis et al. (2017); while rest of the requirements were adopted and interpreted from Chakrabarti et al. (2018).

2.5.2 Customers of the Data Marketplace Platform

Broadly, there can be 2 kinds of customers using the data marketplace platform; namely, ***Data Providers*** who sell data and ***Data Consumers*** who buy the data. Owing to the scope of the focal data marketplace platforms of this research (*many-to-many B2B data marketplaces)*, the customers here comprise only of business organizations who have adopted data-driven business models; but no individual customers.

2.5.2.1 Data Providers

Data Providers are the organizations that publish and sell data on the data marketplace platform. The big data explosion has helped organizations to create business models around the data itself as an offering and reap in economic incentives (Guszcza et al, 2013). The data providers can further consist of 3 kinds of actors:

- ***Data Collectors***: They capture the data either as their main activity (e.g. meteorological measurements, web crawlers etc) or as a biproduct from their main activities (e.g. social media, IoT services etc). They provide raw datasets on the data marketplace platforms.
- ***Data Managers***: These are the organizations that catalogue, clean and parse the raw data into more meaningful and more-interpretable data (Leiponen et al., 2016). They basically perform data curation services like formatting, language translation, identification of outliners etc (van Bommel et al., 2005), and improve the value of the data to be traded on the data marketplace platforms.
- ***Data Aggregators***: These are the organizations that compile data from multiple sources and aggregate to create valued data products. They search, cross reference and

contextualise the data to find correlations or just combine the datasets to create a differentiated data which can be useful for other businesses (Leiponen et al., 2016).

Although these customers perform different key activities, from the perspective of a data marketplace platform, they offer their data on the platform for sale. Hence, they are grouped into one data customer as *data provider.*

2.5.1.2 Data Consumers

Data Consumers are the organizations that search and purchase data on the data marketplace platform. Usually, these are the organizations that have adopted data-driven philosophy in their operations like in their decision making, optimizing business processes or to create data-driven products or data-driven business models (Hartmann et al., 2016). These activities fuelled by the data helps the data consumers understand their customers better, differentiate their offerings to serve them better and thus, attain competitive advantage in their respective markets (Liang et al., 2018).

2.5.3 Functional Components of the Data Marketplace Platform

The functional components here were conceptualised by gaining inspiration from the works of Koutroumpis et al. (2017), Quix et al. (2017) and the *Enterprise Data Marketplace* (EDM) by Wells (2017) which was conceptualised as part of a research conducted by the consulting firm, *Eckerson Group*. During the formulation of these components, the underlying condition which guided the process was that all the conceptualised components should enforce all the functional requirements from subsection 2.4.1 in a comprehensive way. Depending on the platform design of the data marketplace, the object being managed by the data marketplace platform can either be both data and metadata (centralised) or just metadata (decentralised). For simplicity sake, to involve both the platform designs, we use the term "*(meta)data*" to represent the object being managed by the data marketplace platform when discussing the components that have common meaning for both the designs. However, when dealing with specific platform designs, the corresponding term of either *data* or *metadata* is used. The different functional components of the data marketplace platform were formulated as follows.

2.5.3.1 Identity Management

The *Identity Management* is responsible mainly for enforcing the *boundary conditions* for the participants to enter the data marketplace platform and access its services. A screening process can be put in place for the participants to enter the data marketplace platform in order to establish the legitimacy of that participant so that the platform services can be protected from malicious actors. After the entry, the credentials and privileges of the participants must be managed and maintained. To handle these features, 3 services were conceptualised as part of *identity management*; namely, ***induction, authentication*** and ***authorization***. Basically, this component takes care of the security aspects of the data marketplace platform. This component stores and manages the credentials and privileges which can be termed as the identity information of the participants. This identity information can also contain participant profiles with sensitive information like personal identifiable information, payment details et cetera which needs to be protected. Hence, *identity management* should be implemented with utmost secure technologies.

2.5.3.2 Broker Service

Broker Service is the most fundamental component for a data marketplace platform to have as it comprises of the features that reflect the platform aspect of the data marketplaces. *Broker Service* is responsible for 2 kinds of features which are described as follows,

2.5.3.2.1 Backend Features: Data Management Services

The *backend features* comprise of the services which manage the (meta)data which are conceptualised as follows,

- ***Data Cataloguing***: This is a service which involves creating and maintaining the catalogue inventory of every (meta)data present on the data marketplace platform. This service basically showcases the portfolio of the data marketplace.
- ***Data Marketplace Curation***: This service involves 2 activities: *data categorisation* and *data tagging*. Data can be categorised on the high level as *raw data, integrated data* and *aggregated data*. The data can also be categorised based on other context like, by quality, subject area, timeliness, industry etc. Data tagging complements the categories by tagging each data set helping the data consumer to find the relevant (meta)data. Overall, the categories help in arranging the data in a taxonomy helping the data consumers to browse for data while tagging enables data consumers to search for the required data. Curation further involves explicit tagging of the data that is sensitive to privacy, security, legal compliance and other constraints. The activities of data categorisation and data tagging applies for both platform designs as the curation is with respect the data proposition provided by the data marketplace. For a centralised platform, the curation represents for the data that is there on the platform; whereas, for a decentralised platform, the curation represents the data being transacted over the platform by searching and selecting data based on the metadata information.
- ***Data Tracking***: This service tracks the lineage and usage of the transacted data which is appropriately updated on that data's metadata information; thus, enforcing *data provenance*.

2.5.3.2.1 Frontend Features: User Interaction Services

The *frontend features* include the services that provide a marketplace experience for the participants of the data marketplace platform. This basically include features to publish, browse, search, transform and access the (meta)data for the participants. Thereby, they enforce the fundamental platform requirement of matching data consumers to the prospective data providers to fulfil the former's data needs.

Through these features and services, the *broker service* enforces multiple functional requirements *Broker service* enforces the fundamental one, *data exchange platform* specifically through the services of *(meta)data cataloguing, data marketplace curation and user interaction services.* Furthermore, the *broker service* enforces *data governance* through the services of backend features and *data provenance* through *data tracking* service. Overall, by providing the fundamental data marketplace platform services, *broker service* enforces the requirement of *data economy* for the data marketplace ecosystem.

2.5.3.3 Clearing House

Clearing House is the component fundamentally essential for any digital marketplace. The component houses the repository of data exchange transactions information. Every data exchange transaction is recorded and stored in here. This component provides transaction reports essential

for billing, and, help in tracing the lineage of a data product, thus enabling the requirement of *data provenance* and in turn, *data governance*.

2.5.3.4 Data Inventory

Data Inventory is a storage component which reflects the repository of the (meta)data. The broker service orchestrates the processes for the uploading and retrieval of (meta)data from this component. Based on the marketplace platform design, the data can either be stored at the provider site enabling the requirement of *data sovereignty* while the platform housing only the metadata inventory (decentralised platform); or both data and metadata can be housed on the platform (centralised platform). The enforcement of *data sovereignty* is weak on the centralised platform as the data providers participate only based on the intangible trust towards the data marketplace provider i.e. the data provider is expected to trust the marketplace provider's word for what has been done with the data after it left the data provider's premises. However, the metadata maintained in the inventory also contains information about the terms of usage in a contractual form which kind of provides control for the data provider over the usage of his/her data; thus, enforcing *data sovereignty.* Additionally, the component contributes partially towards enforcing *data governance* with the help of the *Broker Service.*

2.5.3.5 Data Exchange Service

Data Exchange Service comprises of the mechanism through which the physical data travels from the data provider to the data consumer in a secure way. Through the defined process, this component enforces the requirement of *secure data exchange*.

2.5.3.6 Data Analytics Service

Data Analysis Service is an additional way of creating value for the participants. We signify this component as the provisioning of data analytic tools which can be used to enrich the datasets into more valuable products. The tools may include data preparation, aggregation, transformation, language translation, visualization and many more. The tools can be provided on the platform in the form of downloadable software or SaaS. These tools are handy for big data players to refine their data offering and make it more attractive in the data marketplace platform. By doing so, these tools can bring in additional revenue to the data marketplace; thus, contributing towards enforcing the requirement of *data economy.*

Using the conceptualisations developed to populate the attributes of functional requirements, customers and the functional components of *HLA framework* as applicable for a generic data marketplace platform, a high-level architecture for the same was built which is illustrated in Figure 4. In addition to the attributes of *HLA framework,* we have also incorporated the dependencies among the attributes which represents *which customer depends on what component for what requirement*. This high-level architecture reflects the answer to the sub-research question, *SQ1* and signifies the ***Pre-MPC Data Marketplace Platform 1.0***.

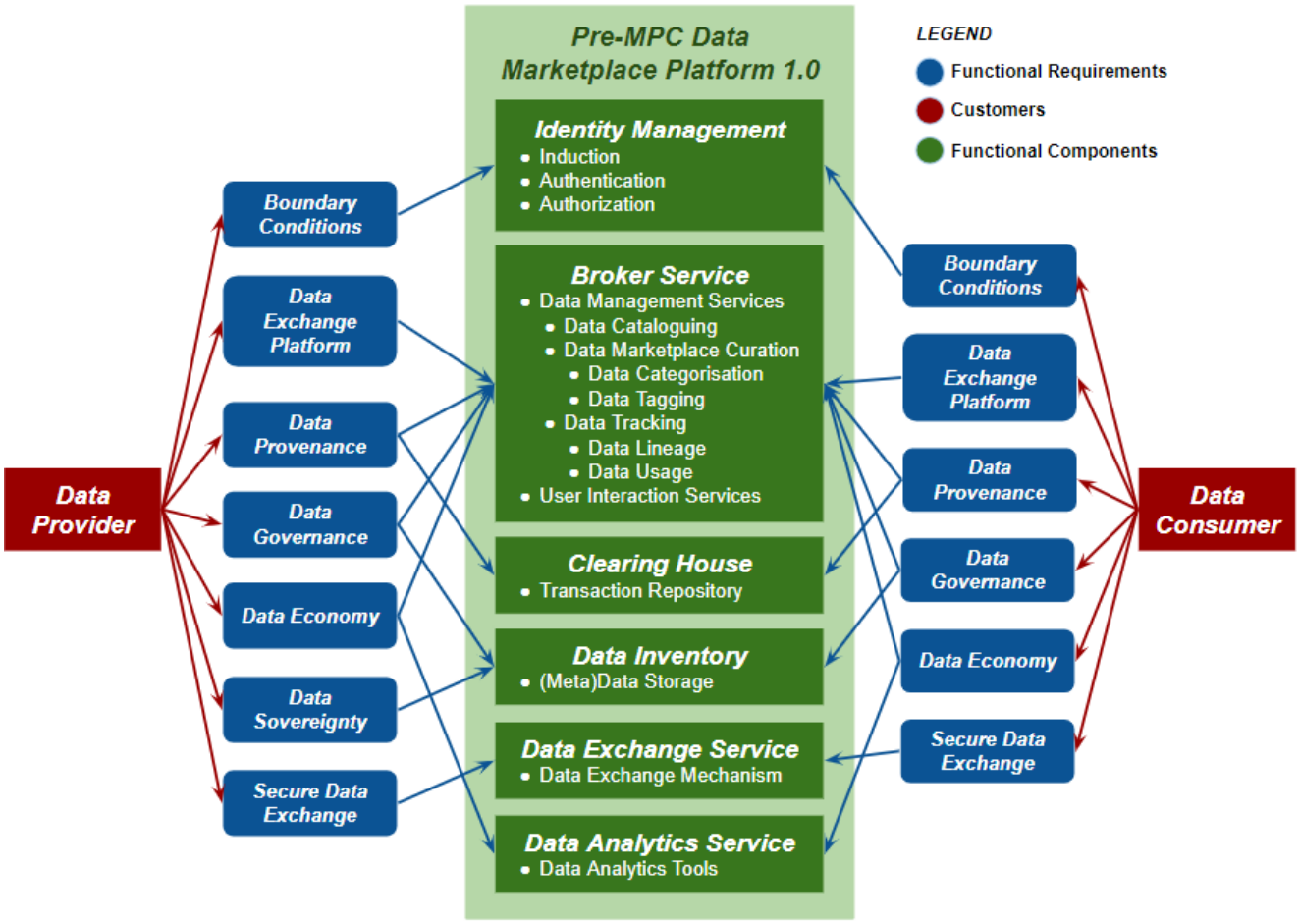


***Figure 4**: High-Level Architecture of a generic Data Marketplace Platform (Pre-MPC Data Marketplace Platform 1.0)*

2.6 Summary

In this chapter, we discussed the research task, *RT1* whose purpose was to establish an architecture for a generic data marketplace platform using just the desk research methods. To accomplish this, the following 2 sub-research questions are answered,

SQ1: How to build an architecture for a generic data marketplace platform?
and
SQ2: How does a generic data marketplace platform look like?

A literature study was conducted on data marketplaces to explore the phenomenon and to understand their fundamental concepts like the definition, different features, relevant actors et cetera. Firstly, the methodology used to search and select the relevant literature on data marketplaces was described. Using the selected literature and their further analysis, we established some of the fundamental concepts associated with the data marketplaces. The challenges associated with the commodification of data was explored and deduced that because of its unique characteristics, data suffers from the issues of *weak protection regime* and *data sharing reluctance.* Then, the phenomenon of the data marketplaces was discussed which involved: dealing with the definition of the data marketplaces; listing the types of data marketplaces followed

by establishing the focal type; and finally, discussing the issues associated with materialising the data marketplaces while describing the different platform designs of the data marketplace platforms as specified by the literature. At this point, a knowledge gap was identified with respect to the comprehensive understanding of the functional aspects of the data marketplaces. It was deduced that the architectural aspects of the data marketplace platforms have never been researched before. Hence, it was decided to build our own architecture for a generic data marketplace platform. Then, the difficulties associated with building a technical architecture with low-level specification for a data marketplace platform were established. Following this, it was decided to build a *high-level architecture* for a generic data marketplace platform reflecting a technological architecture with surface-level (high-level) information of data marketplaces. To carry out this task, *High-Level Architecture (HLA) framework* (Figure 1) was developed consisting of the attributes: *functional requirements, customers* and *functional components*. This framework reflects the answer for the sub-research question, *SQ1* that an architecture can be built for a generic data marketplace platform using this *HLA framework.*

Following this, the *HLA framework* was applied which involved populating the values for the attributes with the information either from the literature or self-conceptualisation as applicable to our focal data marketplace platform. The functional requirements were formalised to be: *boundary conditions, data provenance, data governance, data economy, data sovereingty, secure data exchange* and *data exchange platform.* The customers of the data marketplace platform were idenitifed to be *data providers (data collectors, data managers* and *data aggregators)* and *data consumers.* The functional components were conceptualised to be: *identity management, broker service, clearing house, data inventory, data exchange service* and *data analytics service.* Using these concepts, a *high-level architecture* (Figure 2) was built which reflects a generic data marketplace platform, thus answering the sub-research question, *SQ2;* and signifying the ***Pre-MPC Data Marketplace Platform 1.0.***

We feel the above-developed conceptualisations (*HLA framework* and *High-Level Architecture of a generic Data Marketplace Platform*) to be the answers to the sub-research questions, *SQ1* and *SQ2* respectively because both the answers led us to an architecture; which does reflect a generic data marketplace platform in a comprehensive way. Furthermore, the architecture comprises of all the fundamental elements and functionalities that a data marketplace platform should possess. The limitation here is that the data marketplace platform is represented only at the surface-level (high-level) but not at the technical (low-level) level. For this reason, the resulting high-level architecture only contributes to some extent towards filling the knoweldge gap existing in the literature related to the functional and architectural aspects of the data marketplace platforms but not completely, However, it was already established that this issue does not affect our research objective but helps it by establishing a reduced scope for the subject of data marketplaces in our already mammoth research problem.

3

A Study on Threat Modelling

The first half of the research task, *RT2: To identify the threats associated with the architecture from RT1,* is discussed in this chapter; and the following research question is answered,

SQ3: How to model the threats for the architecture of the data marketplace platform from SQ2?

A literature study was conducted on the process of threat modelling with an underlying criterion to safeguard the fundamental computer security properties: *Confidentiality, Integrity* and *Availability* (***CIA***). The aim of the literature study was to explore and understand the threat modelling process and further, to search for a suitable framework or methodology advocated by different researchers from the academic and non-academic literature. After understanding and comparing the frameworks/methodologies, a new framework was developed to appropriately perform threat modelling on the *Pre-MPC Data Marketplace Platform 1.0* from Chapter 2 which answers *SQ3.*

The rest of the chapter is structured as follows. Section 3.1 describes the methodology used to search and select the relevant literature on the process of threat modelling. Section 3.2 gives an overview on the process of threat modelling by introducing the key concepts and terminology and further relevant concepts related to establishing the context of threat modelling activity. Section 3.3 provides an overview of the threat modelling frameworks and existing threat models. Section 3.4 establishes the context of the threat modelling activity as required for the *Pre-MPC Data Marketplace Platform 1.0*. Section 3.5 introduces the NGCI Apex Classification of Cyber Threat Models which provide a solution with respect to our context. Following this, Section 3.6 describes the new *HLTM framework* designed to be suitable for our context. Section 3.7 summarises the chapter where the focal sub-research question, *SQ3* is formally answered.

3.1 Literature Search and Selection Methodology

The aim of the literature analysis was to determine an approach to carry out threat modelling on the high-level architecture of the data marketplace platform from Chapter 2. Consequently, the focus of the literature search was for a threat modelling methodology which can accommodate a high-level architecture of a technological entity with no low-level technical specification.

With this aim, a simple search was performed on Web of Science and then Scopus with the search phrase, ***"threat modelling"*** which resulted in ***199*** and ***683*** articles respectively. The articles ranged from dealing with threat modelling of specific systems like unmanned autonomous systems; to detecting specific types of cyberattacks like Ransomware; to securing specific domains like cloud, IoT, supply chain environments etc. Clearly, there exists a plenty of literature dealing with threat modelling of a variety of systems. Since studying and comparing each of these methodologies would evidently be a cumbersome job, the strategy was then changed to search for *review/survey* articles which dealt with the analysis and comparison of different threat modelling methodologies. This strategy was expected not only to help in finding a suitable methodology but also in covering bases of threat modelling in different areas, scopes and levels to ensure comprehensiveness of the search. Consequently, a key word search of ***("threat modelling" AND (review OR survey))*** on Web of Science and Scopus yielded ***10*** and ***53*** articles. Out of these, ***2*** articles were identified in the results of Scopus which satisfied our focus to some extent. Firstly, "*Threat modelling – A systematic literature review*" by Xiong & Lagerström (2019) consisting a review of ***54*** articles. Secondly, "*A review of threat modelling and its hybrid approaches to software security testing*" by Omotunde & Ibrahim (2015) comprising of a review of ***101*** articles. The limitation of these articles was that both their review consisted of only software engineering approaches (technical aspects) to threat modelling. Hence, they did not fit our requirement. At this point, we broadened our boundaries of search by conducting the same keyword search, ***("threat modelling" AND (review OR survey))*** on Google Scholar in hoping to find review articles from wider range of sources. This yielded in several results which mostly contained security requirements engineering and security practices. To include the cybersecurity aspect into the search, the key word was refined to ***("cyber* threat modelling" AND (review OR survey))*** given that the threats were investigated with respect to the computer security properties (*CIA*). This resulted in a review article authored by Bodeau, Mccollum, & Fox (2018) as part of The MITRE Corporation working for the Homeland Security Systems Engineering and Development Institute (HSSEDI); which had conducted the survey of threat modelling frameworks; analysed the methodologies and compared them and created a framework out of the knowledge obtained from the reviewed methodologies. Since the article presented comprehensiveness of the phenomenon of threat modelling, it fit our focus of the search in contrast to the review articles found earlier which were limited only to software engineering approaches. The former review article was chosen, and it formed the basis for further literature analysis providing knowledge about different threat modelling frameworks and methodologies. Additionally, secondary literature from the review papers were also analysed as applicable when discussing appropriate aspects. This literature search was conducted till ***20 May 2019***. Hence, any literature published after this date was not considered for this literature study.

Bodeau et al. (2018) uses the term "*cyber threat*" specifically instead of the term "threats". However, the term "threat" involuntarily refers to cyber threats in the realm of technological organizations as they all operate in cyberspace. Even most of the literature (apart from (Bodeau et al., 2018)) use the term "threat modelling" everywhere. The reason can be that since every organization operates in cyberspace lately in some or the other way, every threat can be attributed to being cyber threat either directly or indirectly. Following this reason, we use the term, "threat" throughout the chapter for simplicity, but we evidently mean cyber threats by it as data marketplace platforms are prone predominantly to cyber threats.

3.2 Process of Threat Modelling

Bodeau et al. (2018) define threat modelling as "*the process of developing and applying a representation of adversarial threats (sources, scenarios and specific events) in cyberspace*". Logistically, this process can be carried out in several different ways depending on the context. Microsoft provided a fundamental approach to serve as a starting point for the threat modelling process which was directed towards web applications (Meier et al., 2003). The steps of the process

as developed by Microsoft involved: "*1) Identify security objectives; 2) Create an application overview; 3) Decompose the application; 4) Identify threats; and 5) Identify vulnerabilities*". This approach and its interpretations have been adopted and advocated by many researchers to carry out threat modelling (Steven, 2010; Kamatchi & Ambekar, 2016). EMC added an extra feature to this process in the step of identification of the threats. A library of generic threats was developed to guide the threat modelling activity which simplified the process of identification of threats in EMC's context (Dhillon, 2011). Further, the threat modelling process was adopted in the areas beyond web applications and software development. The process of threat modelling was modified according to the context of the respective areas which led to the advent of different threat modelling frameworks. Currently, the process of threat modelling involves selecting a threat modelling framework and developing a threat model by populating the framework with values as relevant to the intended context (Bodeau et al., 2018). From the populated framework, the threat scenarios which are the representation of the adversarial threats can be constructed and appropriate mitigation controls can be characterised. Since these frameworks and their respective terminology are highly context dependant, the threat modelling process cannot be standardised. This provides a flexibility to design the threat modelling process effectively to the needs of the context and consequently, the resulting threat model would be effectively valid in that context.

It is evident that the crucial aspect of threat modelling process is the formulation of the context in which the threat modelling will be carried out. The following subsections deal with the different aspects of formulation of the context by providing a background on its relevant concepts.

3.2.1 Key Concepts and Terminology

Before diving into the aspects of context formulation, it is important to brush up on key concepts and terminology related to threat modelling. To start off right from the basics, a *model* is defined as "an abstract representation of some domain of human experience, used to structure knowledge; to provide a common language for discussing that knowledge; and to perform analyses in that domain" (Bodeau et al., 2018). The domain here is the threat landscape of cyberspace around the technological organizations.

The terms used in threat modelling involve threat, threat actor, threat vector, threat scenario, attacker, attack, attack vector, malicious cyber activity, intrusion et cetera. These terms are defined differently in different threat modelling approaches based on the assumptions about the context of the technological and operational environment. However, few concepts are generally crucial to be aware of in the threat modelling area. Bodeau et al. (2018) suggests these concepts as,

- undesirable events (***threat or threat event***)
- forces or actors causing the events (***threat source***)
- structured accounts of how the event could cause the harm (***threat scenario***) and
- the resulting harm (***consequence***)

The term *threat/threat event* has different interpretations. The risk assessment guide published by National Institute of Standards and Technology (NIST) in its publication NIST SP 800-30R1 defines *threat* as *"Any circumstance or event with the potential to adversely impact organizational operations (including mission, functions, image, or reputation), organizational assets, individuals, other organizations, or the Nation through an information system via unauthorized access, destruction, disclosure, or modification of information, and/or denial of service"* (NIST, 2012). This definition provides a generic view of threat from a wider scope. A narrower definition from the perspective of the information systems literature is given by The Federal Financial Institutions Examination Council (FFIEC) Information Security Handbook on Risk Assessment (FFIEC, 2016), which reflects our focus of threat modelling, "*Threats are events that could cause harm to the confidentiality, integrity, or*

availability of information or information systems, through unauthorized disclosure, misuse, alteration, or destruction of information or information systems.".

Threat sources comprise of 4 types as identified by NIST SP 800-30R1. They are: *adversarial, accidental, structural* and *environmental*. For our focal system which is a high-level abstraction, *structural sources* are irrelevant as no technical specification is available. The same goes with *environmental sources* as the focal system is a technological platform which is not *directly* affected by *environmental threats*. Although both of these sources come into picture at the further levels of threat modelling. *Accidental* sources are the ones who mean no harm but accidentally take actions that result in harm to the system however, these are dependent on the processes existing in the system which can accidentally go wrong. These are somewhat relevant to our context which will be explicated during the threat modelling activity. Finally, *Adversarial* sources are described as *"individuals, groups or organizations that seek to exploit the organization's dependence on cyber resources (i.e., information in electronic form, information and communications technologies, and the communications and information-handling capabilities provided by those technologies)*" (NIST, 2012). Basically, the *adversarial* sources are the ones with malicious intent who comprise of further aspects, *characteristics* and *behaviours*. *Characteristics* includes further 2 aspects, *capabilities* which reflect the expertise and resources held by the adversaries and *intent* comprising of cyber goals (e.g. gaining access) or intended cyber effects (e.g. denial of service, data breach etc); non-cyber goals (e.g. Financial gain); and risk trade-offs. *Behaviours* are described by *tactics, techniques and procedures (TTPs). "Tactics are high-level descriptions of behaviour, techniques are detailed descriptions of behaviour in the context of a tactic, and procedures are even lower-level, highly detailed descriptions in the context of a technique. TTPs could describe an actor's tendency to use a specific malware variant, order of operations, attack tool, delivery mechanism (e.g., phishing or watering hole attack), or exploit."* (Johnson et al., 2016). The *behaviours* of the adversarial threat agents can be characterised in terms of *threat vector* or *attack vector* they use (Bodeau et al., 2018). *Attack vectors* are *"general approaches to achieve cyber effects, and comprise of cyber, physical or kinetic, social engineering and supply chain attacks*" (Bodeau et al., 2018).

Threat scenario is defined by NIST SP 800-30R1 as "*a set of discrete threat events, associated with a specific threat source or multiple threat sources, partially ordered in time*" (NIST, 2012). This relates to the 7 stages of hacking suggested by D. A. Smith (2017) where each stages signifies a single threat event with the whole affair translating to threat scenario.

And finally, *consequences* are the harm caused in terms of effects on information and information systems. The *cyber effects* are expressed as loss of confidentiality, integrity and availability and are translated into effects on the systems, business functions, organization and its customers.

These are some of the key concepts and terminology which are relevant to the threat modelling activity. In the coming subsections, the different aspects of formulation of the context of threat modelling activity are discussed which would later guide us to formulate the context of our threat modelling activity.

3.2.2 Scope of Threat Modelling

Bodeau et al. (2018) identified different scopes at which threat modelling can be performed. The scope within an organization ranges from *information system tier* (implementation/operations level), expanding to the *mission/business function tier* (business/process level) and then to the higher-most *organizational tier* (executive level). Beyond the confines of organization, 2 additional levels also apply which are: sector, region or community-of-interest (COI) level and national or transnational level (Bodeau & Graubart, 2014). These levels are illustrated in Figure 5. The significance of threat modelling at each level is as follows,

- ***Information system tier***: Threat modelling at the level of implementation or operations of an information system can motivate the design decisions and the selection of security controls in different stages of System/Software Development Life Cycle (Bodeau et al., 2018).
- ***Mission/Business function tier***: Threat modelling at this level can influence different aspects of business process architecture, enterprise architecture and the information security architecture for the business function in focus (Bodeau et al., 2018).
- ***Organizational Tier***: At this level, threat modelling is carried out to evaluate the threat environment in which the organization is operating. This means the ecosystem consisting of the suppliers is also accounted in the threat modelling activity (Bodeau et al., 2018). This leads to the development of shared threat models which accounts for different actors in the ecosystem.
- ***Beyond the levels of organization***, threat modelling serves the research agenda promoting threat information sharing between organizations, nations et cetera. This helps in improving the security intelligence scene at the sector or the national level.

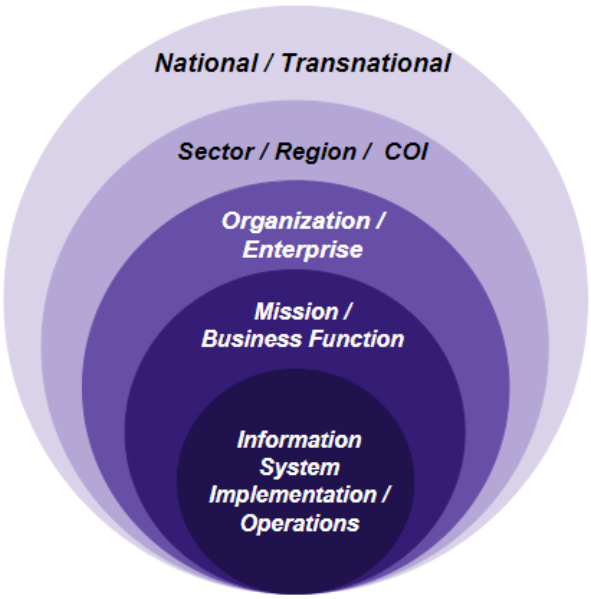


Figure 5: *Scope of Threat Modelling*
Source: Bodeau et al. (2018)

3.2.3 Approach of Threat Modelling

Bodeau et al. (2018) propose 3 approaches for the threat modelling activity as follows,

- ***Threat Centric***: A known threat is coupled against the focal system to understand its effect on the system which guides the security development of that system.
- ***System Centric***: Here, the systems subjected to threat modelling are modelled first specifying their architecture and boundaries in the context and then identifying the relevant threats to the modelled system.

- ***Asset Centric***: The assets of value in the context which are sensitive to the threats are identified and then the threats are characterised that could reach and affect the assets.

The approach chosen gives a focus for only one aspect as a starting point, but assumptions must be made about different aspects to determine the scope of the primary aspect. *Figure 6* illustrates the approaches and different aspects which come into picture with respect to each approach.

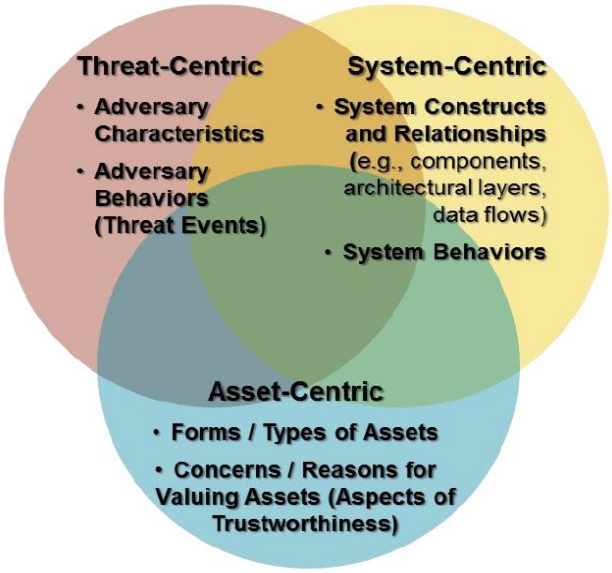


Figure 6: *Threat Modelling Approaches*
Source: Bodeau et al., 2018)

3.2.4 Purpose of Threat Modelling

Apart from the *scope* and *approach*, another dimension which is relevant to threat modelling is the *purpose* served by the threat model. Bodeau et al. (2018) suggests the following purposes which are fulfilled by employing threat modelling.

- ***Risk Management:*** Risk management, as conceptualised by NIST, possesses 4 component processes: "*risk framing, risk assessment, risk response and risk monitoring*" (NIST, 2011). These stages respectively involve:
  - forming the assumptions of the context and formulating the threats in that context;
  - judging the severity of the identified threats in the environment and determining the likelihood and consequences of the threats;
  - adopting appropriate mitigation controls for the identified threats; and
  - designing the security architecture to prevent these threat events from occurring (Bodeau et al., 2018).

  Threat modelling is a part of risk framing where the threats are identified. The threats are represented either in the form of generic threat events or specific threat scenarios based on the level of detail expected. The threat model along with the risk management helps in developing the security portfolio of the organization.

- ***Cyber Wargaming: "****Cyber wargaming is a method of exercising and examining, in a modelled environment, human performance and decision-making or system characteristics and*

outcomes in the context of a cyber attack scenario" (Bodeau et al., 2018). The cyberattack scenarios are generated with the help of threat modelling.

- ***Technology Profiling and Foraging:*** The identification of threat events and threat scenarios through threat modelling can support the evaluation of capabilities of existing security controls, practices and technologies (profiling) and even scouting for the technologies of potential interest (foraging).
- ***Systems Security Engineering / System Design Analysis:*** In this scenario, the threat modelling supports the System Development Life Cycle (SDLC) comprising of requirements definition, analysis and design, implementation, testing and operations and maintenance. Ultimately, threat modelling enables the designing a secure system instead of adopting security controls after the design of the system.
- ***Security Operations Analysis:*** This purpose addresses the post-design security formulation of the system which is referred as cyber defending. This involves a proactive threat modelling approach including activities like threat hunting, continuous monitoring & security assessment and DevOps. This purpose can be effectively fulfilled by threat information sharing which helps for the overall security situation to grow.

The process and the elements of the threat modelling can be oriented according to the needs by formulated the context comprising: scope, approach and the purpose of the threat modelling activity. The conceptualisation of the scope is illustrated in Figure 7.

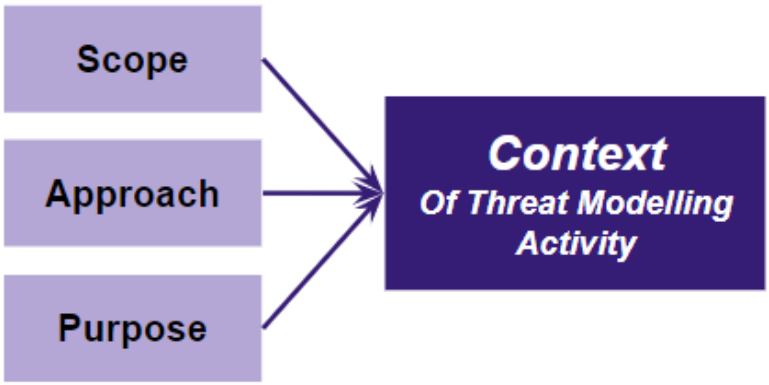


***Figure 7**: Conceptualisation for the Context of Threat Modelling*

3.3 Threat Modelling Frameworks

To gain more insight towards framing the context for our threat modelling activity, threat modelling frameworks which operate in different contexts were reviewed. Similar to Bodeau et al. (2018), the frameworks were categorised for the discussion ahead based on their purpose; i.e. for *Cyber Risk Management*, for *System Design & Analysis* and for *Threat Information Sharing*. A widely-used methodology in each category and later a few populated threat models which contain commonly identified threats already familiar are discussed.

3.3.1 Frameworks for Cyber Risk Management

There are several frameworks which help the purpose of cyber risk management. One such approach was developed by National Institute for Standards and Technology (NIST) in their various publications which contain threat modelling as an explicit part of their risk management process. As defined earlier, NIST's risk management framework contains 4 components: *risk framing, risk assessment, risk response and risk monitoring*. Thread modelling is part of their first component, risk framing. They define risk framing in their publication NIST SP 800-39 as, "*the set of assumptions, constraints, risk tolerances, and priorities/trade-offs that shape an organization's approach for managing risk*" (NIST, 2011). This step also involves assuming about the threat environment of the focal entity. The threat environment here is described as threat sources and threat events including the types of adversarial TTPs and adversarial characteristics (capabilities, intent etc). These assumptions form the threat model and the risk assessment helps in prioritising the threats and documenting them for the next step, risk response. The threat model is updated every time the risk assessment is carried out. They provide a representation threat model which comprise of; a taxonomy of threat sources with their characteristics, a set of threat events and a taxonomy of predisposing conditions which help in judging the likelihood of the threats. This initial threat model forms the starting point to start the brainstorming of the assumptions of the focal entity's context to develop its threat model. Bodeau et al. (2018) have surveyed several other frameworks and methodologies dealing with cyber risk management. Their work can be referred for more detailed analysis and relevance of the frameworks.

3.3.2 Threat Modelling for System Design and Analysis

This category contains a plenty of highly structured threat modelling approaches which supports the system design decisions and its development process. The survey article by Xiong & Lagerström (2019) as mentioned earlier consists of the analysis of 54 articles employing different methodologies which only deal with the purpose of system design and testing. Bodeau et al. (2018) also have reviewed few methodologies out of which the most popular one is reviewed here; the widely referred methodology developed by Microsoft as part of their secure Software Development Life Cycle (SDLC) agenda, ***STRIDE*** model.

STRIDE is an acronym which stands for "*Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service and Elevation of Privilege*" which represent the general categories of threat vectors applicable in software environment. STRIDE primarily helps in the steps of *threat identification* of the threat modelling process proposed by Microsoft. It is flexible and highly dependent on the system specification and architecture. Each of the components and their interaction with each other and the flow of data are analysed, and STRIDE mnemonics are applied to each component to identify threats specific to that component. Based on these findings, the developer can identify different bugs in the system and decide how to fix them. STRIDE is helpful in identifying threats in the system but further techniques like threat trees, attack trees etc are needed to model the threat events and scenarios. The STRIDE model is like the risk framing step of the NIST framework but in a software environment. It is supported by another model called ***DREAD*** (Damage, Reliability, Exploitability, Affected Users and Discoverability) which is also developed by Microsoft to evaluate the threats and choose the relevant threats to mitigate; like risk assessment step of NIST framework. Several researchers have used and have recommended STRIDE framework to model threats in a variety of environments by customizing it to fit their requirements (Steven, 2010; Kamatchi & Ambekar, 2016; Marback, Do, He, Kondamarri, & Xu, 2013). It is important to notice that STRIDE takes *system centric* approach to model threats for the purpose of system design and testing.

A different way of system-centric threat modelling is proposed by Uzunov & Fernandez (2014) in which they decompose the system architecture into its generic functional components and

develop a taxonomy of threats based on the characteristics of each component. The taxonomy is used as a reference when the threat assessment is carried out for specific systems and the newly identified threats are updated in the taxonomy. This is the most applicable way of doing system-centric threat modelling, but it requires complete specification of the system and an expert threat modeller. There are many more methodologies which take different approaches of threat modelling to system design. For example, Intel's Threat Agent Risk Assessment (TARA) which takes the threat-centric approach (Rosenquist, 2009) and IDDIL/ATC methodology which takes an asset-centric approach with the first step being to identify and characterise the assets in the context (Muckin & Fitch, 2017).

3.3.3 Threat Models for Threat Information Sharing

The threat modelling frameworks discussed so far can direct the process towards developing the threat models. They were used in the initial years when threat modelling was an infant field to research. But since then, the field has evolved, and more sophisticated techniques have been developed to carry out threat modelling. As a result, the previously discussed frameworks are often not used in the organizations. From the representation threat model of NIST SP 800-30R1, organizations develop hybrid or customised approaches for various purposes suited to their business processes. Here, some of the threat models are discussed which were developed for various purposes but help by lending the information about a variety of techniques used by threat actors in different environments. Bodeau et al. (2018) identified a few threat models which include 2 kinds: enterprise-neutral and enterprise-oriented threat models.

3.3.3.1 Enterprise-Neutral Threat Models

Enterprise-neutral threat models consist of adversary characteristics and behaviours consisting of attack techniques within a general technological environment. The focus here is only on the threat event with adversary techniques and do not incorporate information about enterprise characteristics like its architecture, assets and systems. Basically, they take a threat-centric approach. Some of the examples include ATT&CK (Adversarial Tactics, Techniques & Common Knowledge), CAPEC (Common Attack Pattern Enumeration and Classification), OWASP (Open Web Application Security Project) etc.

ATT&CK is developed by the MITRE Corporation (The MITRE Corporation, 2015) and provides an account of adversary behaviour within an enterprise network i.e. post-access through a successful entry exploit (Bodeau et al., 2018). ATT&CK consists of a repository of adversary attack techniques which operate in a network powered by Microsoft Windows environment. The repository consists of 10 categories of tactics with each tactic containing a list of attack techniques and potential mitigations. The tactic categories are: persistence, privilege escalation, defence evasion, credential access, discovery, lateral movement, execution, collection, exfiltration and command & control. Like ATT&CK, CAPEC model provides a catalogue of attack patterns with more detail than ATT&CK which help in categorising the attacks in a meaningful way; OWASP comprises of 12 categories of attacks applicable in web applications. These models lend several categories of adversary TTPs which can be used to model the threats in the realm of the focal context.

3.3.3.2 Enterprise-Oriented Threat Models

These are the threat models generated after the threat assessment of particular enterprises. Since the models contain sensitive enterprise-specific information about the ways it could be attacked, these are generally not shared. However, Bodeau et al. (2018) identify 3 generic models in this

category which indirectly deal with enterprise-specific threat modelling. One of these models is MITRE's Threat Assessment and Remediation Analysis (TARA) which will be discussed here in brief.

MITRE's TARA is actually a methodology developed for identifying threats to a system and determine countermeasures (Wynn, 2014). The threat identifying component of the MITRE's TARA is called Cyber Threat Susceptibility Analysis (CTSA) which identifies and evaluates potential cyberattack events and patterns. Like the previously discussed populated threat models, CTSA also builds a threat catalogue focussing on the attack vectors. Additionally, MITRE's TARA contains a taxonomy of vector groups and a set of tools which map the attack vectors to different system environments and technologies. This is the differentiating feature of MITRE's TARA compared to other methodologies and otherwise, MITRE's TARA proposes a threat modelling process like that of NIST SP 800-30 and Microsoft: identify the scope, architecture and technological components; make assumptions about the types of adversaries and techniques; and identify the threats appropriate to the scope and assumptions. MITRE's TARA also has its own way of assessing and prioritising threats to mitigate. To sum up, MITRE's TARA, it is useful for threat information sharing as it contains a catalogue of attack vectors and tools to map them to the system environments.

3.3.4 Reflection on the Frameworks

Firstly, the risk management framework developed by NIST in their publications are very generic. The method is highly flexible and depends on the threat modeller to define the detailed tasks as the framework just motivates the threat modeller with relevant aspects; but the methodology does not direct him/her with detailed tasks. Hence, this framework can act as a starting point to learn different aspects of threat modelling, but the threat modeller should be aware of the detailed information of the scope in which he is operating to form concrete assumptions to start the threat modelling process. However, the catalogue of different taxonomies helps in the step of threat identification. Since the methodology is generic and involves lot of assuming and conceptualising, the threat modelling can be time consuming and a tedious process; and needs managers and technicians working together. Then, we discussed STRIDE. It takes a software engineering and system centric approach. Evidently, it can only be carried out by an expert technician and the manager has a lowest role to play in the activity. Although its categories are high-level, it provides a starting direction to decide on security aspects of system design. The reflection here is like that of NIST with one exception that STRIDE applies only in the scope of information systems. On the other hand, the threat modelling methodology of Uzunov & Fernandez (2014) is specific and provides concrete steps to carry out the threat modelling which makes the methodology straight forward; but the framework applies specifically to the distributed networks and also it needs a technically expert threat modeller. Table 5 lists the above-discussed threat modelling frameworks and threat models with the characteristics of their respective contexts.

***Table 5**: Reflections on Different Threat Modelling Frameworks*

Framework	Scope	Approach	Purpose	Technical Expert needed?
NIST SP 800 30R1	Organization, Mission, System	Flexible – Can be made Threat, System or Asset centric based on the information available	Risk Management	Technical Expert translates to more Validity
STRIDE	System	System-Centric	System Design Analysis	Depends on the context specification

Uzunov & Fernandez (2014)	System (Distributed networks only)	System-Centric	System Design Analysis	Yes
Intel's TARA	Organization, Mission, System	Threat-Centric	System Design Analysis	No
IDDIL/ATC	System	Asset-Centric	Risk Management	No
ATT&CK	System (post network entry)	Threat-Centric	Threat Information sharing	Depends on the context specification
CAPEC	System	Threat Centric	Threat Information Sharing	Depends on the context specification
OWASP	System (Web applications only)	Threat-Centric	Threat Information Sharing	Depends on the context specification
MITRE's TARA	Organization, Mission, System	Partly System-centric and partly Threat-centric	Risk Management and Threat Information Sharing	Yes, but a lesser expert compared to other expert methods

From the above reflection, it is understood that the level of detail in which the context is described, dictates the specificity of the threat modelling process. For a context described in great detail, a specific threat modelling process can be tailored which would be highly-structured and highly effective within the context because of which relatively less expertise is needed as the validity is guaranteed by the process itself. Furthermore, the duration of the process depends on the context. Some of the examples here include: Uzunov & Fernandez (2014), Socio-technical Framework by Sabbagh & Kowalski (2015) etc. On the other hand, if the context is not available in detail, then the validity and the time duration depend on the expertise of the threat modeller who needs to make informed assumptions about the context to carry out the threat modelling with generic methodologies. The examples here are: NIST, STRIDE, MITRE's TARA, IDDIL/ATC etc.

Ultimately, we can deduce that the process of threat modelling is highly dependent on the context and to what extent of detail it is available and described. More detailed the description of the context, more effective is the threat modelling process and more valid is the developed threat model. The availability of different varieties of focal entities and their varying contexts there cannot be a single comprehensive framework for threat modelling of all the contexts.

3.4 Context of our Threat Modelling Activity

Based on the understanding of the different threat modelling frameworks and threat models, the context of the threat modelling for the *Pre-MPC Data Marketplace Platform 1.0* from Chapter 2 was formulated and is presented here by specifying it in the language of threat modelling: scope, approach and purpose.

Firstly, the ***scope*** was established. Since, the focal entity, the data marketplace platform is a technological platform represented with a high-level architecture, the corresponding functional components can be considered as individual information systems which can be implemented with technology alone without any human actor needed. However, there is not technical specification of these information systems but on the contrary, only features are specified which translate to the

business functions of those information systems. Hence, the scope of the threat modelling activity was formalised to be at the level of ***business functions***.

Furthermore, the kind of threats and the detail in which the threats are described should be established. As mentioned earlier a threat can be described in 3 levels of detail namely, tactics (high-level), techniques (medium) and procedures (low-level); all of which are represented through attack vectors which can be described appropriately at 3 levels. Since, the unit of analysis here is only the technological components with just business functions and no technical specification, the threats were decided to be described with attack vectors (cyberattacks) at a high-level; which reflect the adversarial dimension of threat source. Related to the other threat sources, they are included when they apply during the threat modelling process. The principle behind the kind of threats is to find the cyberattack vectors applicable which are described later at high-level (*tactics*). For example, DDOS attack on a Server signifies a *high-level* description of threat while the whole logistics of that DDOS attack used on a specific server which entails every step involved in the attack process reflects a *low-level* description of the threats. This implies that the threat modelling activity could be performed by managerial level expert with no need for technical experts.

The ***approach***, as we have explained it, should comprise of the information we know about the focal entity, and the rest of the aspects are to be assumed in the threat modelling process. On these lines, the absence of the technical specification eliminates the system-oriented approach while the already decided specification of the tactic-level (high-level) handling of threats eliminates the threat-oriented approach. The information we do know about the data marketplace platform is with respect to the functional components and their business functions. Consequently, the assets associated with those business functions can be modelled first which can later drive the whole threat modelling activity which entails making assumptions on system and threat ends. Hence, the approach was decided to be ***asset-centric***.

Choosing the ***purpose*** was a straight forward one the aim is to identify the threats associated with the data marketplace platform to understand the threat landscape of the data marketplace platforms, in other words to analyse the risk associated with the data marketplace platforms. Hence, our purpose was on the lines of ***risk framing*** step of *risk management.*

3.4.1 Implication of the Context Formulation

Essentially, the context can be represented by a single statement as, ***"to establish the assets associated with the business functions of each functional component of the high-level architecture of a technological entity and later, assume a system specification on which applicable cyberattack vectors (described at a high-level) can be identified"***. As seen in Table xx, there was only one framework which satisfies our context, *NIST SP 800 30R1.* However, as established already in our reflection (section 3.3.4), this framework entails the necessity of a technical expert who can make credible assumptions about the assets in the functional components, such that a valid threat model can be generated. We do not possess this expertise as we are not technical individuals. The other option is to build a valid threat model is to have a *specific* framework applicable to the threat modelling activity for our context. There is no such framework as most of the threat modelling literature is directed either towards the software engineering area (technical) or the ones whose context are specified in a great detail to the level of infrastructure and practices in an organization. As a result, there exists a gap in the literature related to the threat modelling at the scope of business functions for the technological entities.

3.5 NGCI Apex Classification of Cyber Threat Models

Bodeau et al. (2018) addresses a problem that there is no threat modelling framework or methodology which could comply to all the contexts. He further stresses on the need for a threat modelling framework which can be customised to different purposes and used at multiple levels and scales. For this agenda, Bodeau et al. (2018) conceptualised a classification containing threat models in which the threats are described at all the levels in respective threat models (*tactics, techniques and procedures*). The classification was done as part of their NGCI Apex Program and it contains 3 threat models: ***High-Level Threat Models, Detailed Threat Models*** and ***Instantiated Threat Models.*** They are described as follows and the kind of threats dealt in each threat model is illustrated in Figure 8.,

- ***High-level Threat Models:*** These contain threat events described in general terms which support high-level or sector wide risk assessment, cyber wargames or technology profiling and foraging.
- ***Detailed Threat Models:*** These support technology evaluation in which threat events are described with a little more detail in terms of specific systems, technologies or targets.
- ***Instantiated Threat Models:*** These are low-level threat models containing detailed threat scenarios which help in developing detailed cyber playbooks. These models are dependent on the system architecture and hence, these models are usually developed by the organizations themselves and are not shared to the external entities like academia since they contain sensitive information.

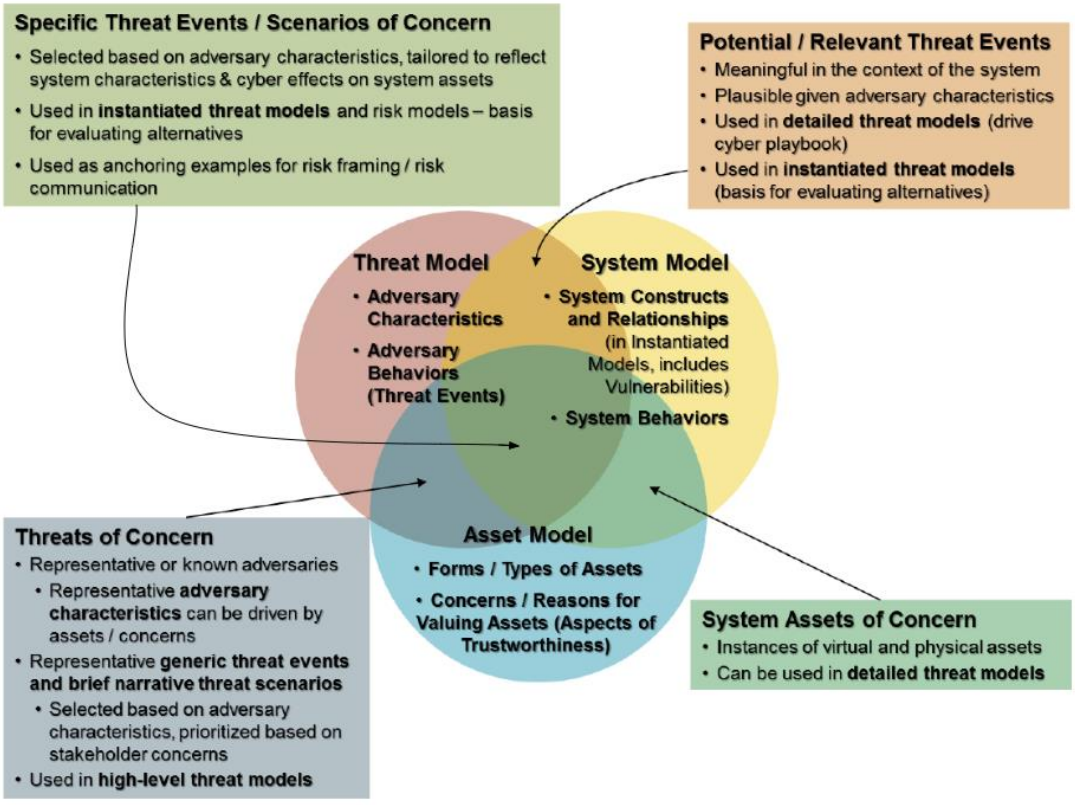


Figure 8*: Types of threats in the Threat Models of NGCI Apex Program*
Source: Bodeau et al. (2018)

The block of ***Threats of Concern*** in Figure xx suggests generic threat events, brief narrative threat scenarios and adversary characteristics which are driven by assets. This interpretation of threats

related to the kind of threats we formulated to find in section 3.4.1. Hence, the conceptualisation of the *High-Level Threat Model*, *comprising of cyberattack vectors described on a high-level which affects the assets,* was considered as the reference to build our threat model as it relates to our context.

3.6 HLTM Framework

We decided to design a framework of our own as applicable to our context which is, *to establish the assets associated with the business functions of each functional component of the high-level architecture of a technological entity and later, assume a system specification on which applicable cyberattack vectors (described at a high-level) can be identified*. We rephrase this context into "***performing high-level threat modelling of the high-level architectures of the technological entity***" because it is driven by the concept of *high-level threat models* from the NGCI Apex Classification and the focal system, which is the high-level architecture of any technological entity.

Consequently, the framework was named as ***High-Level Threat Modelling (HLTM) Framework*** The philosophy of the framework is to break down the focal high-level architecture of the focal technological entity into its functional components, identify the business functions associated with the components and identify the threats which affect those identified business functions. The framework gives a simple structure to identifying the *high-level threats of concern* to the technological entities. Essentially, to represent in the language of threat modelling, the framework operates in the context of asset-centric (*approach*) threat modelling of the business functions (*scope*) to deduce the risk (*purpose*) associated with the focal entity. The detailed description of the context is already established in section 3.4. The framework consists of 6 constructs: ***Functional Component, Business Function, Threat, Cyber Effect, Business Consequence*** and ***Mitigation Technique.*** These are described in the following subsections.

3.6.1 Functional Component and Business Function

Both the constructs, functional component and business function, constitute the *asset* dimension of the framework. An asset is an entity which is a constituent of the system responsible for its value. In the information systems environment, asset can be defined as "*any data, device or other component that supports information-related activities, which can be illicitly accessed, used, disclosed, altered, destroyed and/or stolen result in loss*" (Jones, 2005). The choice of the assets for consideration in the threat modelling process is dictated by the business functions associated with each functional component in the high-level architecture.

Firstly, the functional components are mapped to their respective business functions. A functional component can be responsible for multiple business functions. Then, for each business function, a basic system asset is assumed which fundamentally enables the respective business function. This step signifies assuming of the *system* dimension of the threat modelling activity. The mapping gives a baseline of the low-level technical specification for which threats could be identified.

In our case of technological platform, the assets can be attributed to IT systems. The IT system assets can have different characteristics. The Single European Sky ATM Research (SESAR) differentiates 2 categories of IT assets in their Security Risk Assessment Methodology: Primary and Supporting assets (Marotta et al, 2013). They characterise primary assets as the intangible functions, information, processes, services and activities. Supporting assets are the tangible systems or components which contain vulnerabilities through which a threat agent can attack and compromise the primary asset; for example, websites, communication channels, database, server

etc. We incorporated this concept of IT assets (*primary asset* and *supporting asset*) for assuming the assets associated with the business functions.

3.6.2 Threat

By threats in this framework, as mentioned in section 3.4.1, we refer to cyberattack vectors which are just mentioned at high-level instead of describing the logistical process of the cyberattack happening to the assumed IT system asset. The cyber threats generally revolve around the vulnerabilities in the system and the cyberattacks which take advantage of the said vulnerabilities. Since the system under evaluation is a high-level architecture with no technical specification, the vulnerabilities are excluded from the framework. Instead, each IT asset is considered for each business function and the cyberattack vectors appropriate to the focal IT asset is identified and attributed as its threat.

A cyberattack on a broader perspective, generally consists of 7 steps which are listed in Table 6 (D. A. Smith, 2017). Each of these steps can involve intermediate attacks which form the building blocks to a broader cyberattack. These cyberattacks are listed in Table 7 which form the representative values that can be used during threat modelling. The list is not exhaustive and other cyberattacks can also be included for threat modelling appropriately.

***Table 6**: 7 steps of a Cyber Attack*
Source: D. A. Smith, (2017)

Steps	Description
Reconnaissance	Before a full-fledged cyberattack, the attacker identifies a target and explores the information related to the target.
Scanning	After the identification of the target, the attacker searches for vulnerabilities by scanning the systems through attacks like resource enumeration and browsing (Table xx).
Access and Escalation	Once the weak spot is identified, then attacker tries to gain access to the system and then escalate the privileges to move freely with the system environment. Ex: Password attacks
Exfiltration	The attacker now attempts to access sensitive assets like data and tries to extract it. Ex: Storage attacks
Sustainment	The attacker seeks to remain undetected and have unrestricted access by installing malicious programs like root kits which allows the attacker to return as and when desired.
Assault	Now, the attacker can sabotage the system either by modifying the system or disrupt it entirely by disabling it. This means the attacker has full control of the system and it is too late to defend it.
Obfuscation	This step happens when the attacker leaves a signature behind in the system to brag about his/her conquests. This usually involves confusing or diverting forensic investigation through log cleaners, spoofing, misinformation, zombie accounts, trojan commands etc

***Table 7**: General Cyber Threats to IT systems*

Cyber Threat	Description
Botnet	A botnet is a network of remotely controlled machines used to launch wide-scale denial of service attacks against specifically targeted resources (Zhang et al., 2011).

Denial of Service (DoS, DDoS)	A Denial of Service attack consists an attempt to impeach users from accessing data or services provided by an information system (Zlomislic et al, 2014).
Eavesdropping/ Traffic Analysis	This is a form of attack where the attacker attempts to capture and analyse network data packets in the communication channel in order to identify any information that may be relevant for other types of attacks.(Fu, 2005)
Injection attacks	This attack refers to a broad class of attack vectors through which the attacker injects malicious input to a program. Particularly, SQL injection attack is considered very dangerous as the attacker can gain access to the database with sensitive data by injecting malicious value at the input field (Muscat, 2019).
Malicious code/ Payload	This is a generic family of attacks all of which involve harmful code or script designed to be executed by programs, operating systems, web servers, and any other IT device, resulting in undesired effects. These are usually carried by viruses or worms (Al-Mohannadi et al., 2016)
Man-in-the-Middle	This form of attack is a specific case in the eavesdropping type of attacks, in which the attacker interposes between the sender and the receiver and misleading them into believing their communication line is direct and secure This allows to either intercept confidential information or altering it unknowingly to the legitimate communication participants. This attack affects the confidentiality and integrity of the data in the communication channel (Conti et al., 2016).
Password attacks (Brute-force, Dictionary, Cookie Replay)	In this form of attack, the attacker attempts to identify a password or an encryption key through exhaustive checks or through cookie information from the browser until the correct string is identified (Hansman & Hunt, 2005).
Resource enumeration and browsing	This is a type of attack through which the threat actor is able to obtain from a targeted system the list of the resources that are present in the system, therefore enabling the threat actor to refine the targeting process of such resources and their consequent browsing (OWASP, 2018).
Malware/Viruses	Viruses and malware are types of malicious code/payload with various objectives, among which can be mentioned replication, data manipulation or destruction etc (Bishop, 1991)

3.6.4 CIA Violated?

Information security objectives are represented on a high level with the triad of computer security properties – CIA*: Confidentiality, Integrity* and *Availability*. We use the same triad to represent the computer security objective violated for the focal IT asset by each cyber threat identified in the previous step. The properties are described as follows,

- ***Confidentiality***: The property that the information and the services should be made available to only authorised individuals, entities or processes.
- ***Integrity***: The property of safeguarding the accuracy and completeness of information assets.
- ***Availability***: The property of information assets to be accessible and usable upon demanded by an authorised entity.

Each threat to the IT system assets results in a degradation to one or more of these properties.

3.6.5 Business Consequence

Business Consequence construct describes the adverse effect caused by the threat to the focal business function under consideration or to the whole technological entity in general. This construct helps in representing the adverse effects of the threats in the language of the business aspects, as opposed to the computer security properties (CIA) mentioned in the previous step. The business consequence construct can have variety of values ranging from financial loss, reputational loss, functional loss, regulatory impacts or environmental loss. Although, the value to be filled here is highly dependent on the business function under consideration.

3.6.6 Mitigation Technique

This construct completes the circle of the whole threat modelling activity by recommending the appropriate security techniques which can mitigate the identified threats. The mitigation techniques can comprise of concrete mitigation technologies, protocols, policies and security procedures. The common security controls used are listed on a high-level by (Northcutt, 2018) in his white paper published as part of research at SANS institute. These are: *Security Awareness Training, Firewall, Anti-Virus, Intrusion Prevention System, System Monitoring, Intrusion Detection System and Encryption*. In addition to these, any other techniques and to any extent of detail can also be used to populate this construct.

3.6.7 Reflection on the HLTM Framework

The framework satisfies threat modelling context by the constructs, *Functional Component* and *Business Function* constitute the ***asset*** dimension; the assumption of basic IT system asset specification reflects the ***system*** dimension; and finally, *Threats, Cyber Effect & Business Consequence* constitute the *threat* dimension.

Furthermore, it can be deduced that all the threats and their respective business consequences to each business function reflects the high-level overview of the threat landscape around the focal technological entity. We termed this conceptualisation as ***High-Level Threat Landscape*** of the focal technological entity owing to the *high-level philosophy* dealt so far. The conceptualisation and the resulting HLTM Framework are illustrated in Figure 9 and Figure 10 respectively.

***Figure 9**: Conceptualisation for the Threat Landscape*

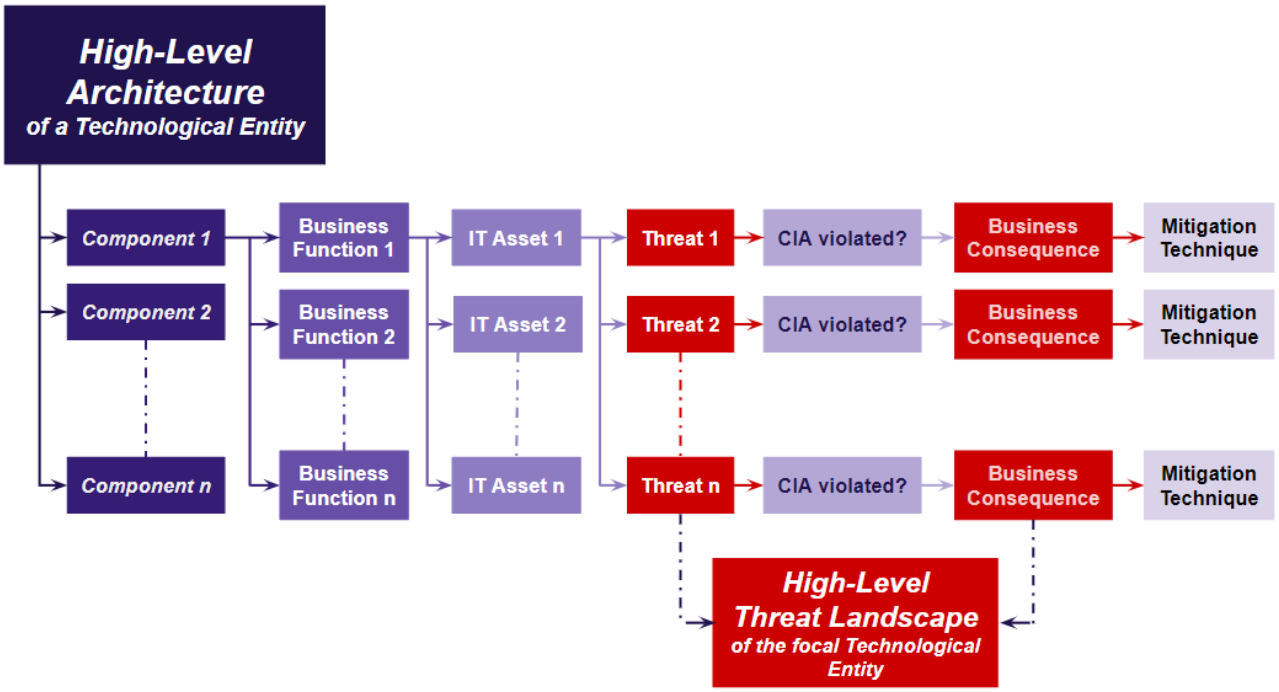


***Figure 10**: High-level Threat Modelling (HLTM) Framework*

Because of this conceptualisation, the framework was deemed fit to be applied on the high-level architecture of the data marketplace platform from Chapter 2 as it could result in an overview of the threat landscape of the *Pre-MPC Data Marketplace Platform 1.0.*

However, because of the mapping of business functions to the basic IT assets i.e. the baseline low-level technical specification, the resulting threats and their respective business consequences represent only the *baseline threat landscape of the focal technological entity*. This can be attributed as a limitation of the HLTM framework. This situation can be improved with multiple iterations of threat modelling as and when more knowledge is learnt on the low-level technical specification of the focal technological entity; which would further result in the *low-level threat landscape*.

Apart from our research objective, the *HLTM framework* is a valuable addition to the family of threat modelling frameworks as there is none existing to address the context it is operating in; which is *to perform high-level threat modelling for the high-level architectures of the technological entities*. All the constructs in the framework are operating according to the *high-level philosophy* with almost no technical specification of the system required. Hence, the identification of the threats can be done even by a manager. However, the framework demands some basic level of technical expertise of cybersecurity which comes handy during the application of the framework like, assuming the supporting assets in IT asset stage, knowledge of which cyber threats could affect what kind of systems et cetera. Essentially, greater the technical expertise of the threat modeller, higher is the validity of the threat model. In that case, threat modelling by a technical expert results in a more valid threat model. In addition to this, the framework can be used perform low-level threat modelling of specific technical architecture of technological entities. The only change would be to instead of assuming the IT assets the constituents of the available low-level architecture can be filled in the IT asset construct. Hence, the framework is flexible enough to adopt between high-level and low-level threat modelling.

3.7 Summary

In this chapter, we discussed the first half of the research task, *RT2, the* purpose of which was to figure out a methodology to identify the threats associated with the *Pre-MPC Data Marketplace Platform 1.0* and answered the following sub-research question,

SQ3: How to model the threats for the architecture of the data marketplace platform from SQ2?

A literature study was conducted on threat modelling to understanding the process of the threat modelling and its subsequent fundamental concepts and requirements. On searching for literature, we found a survey article (Bodeau et al., 2018) which dealt with the topic of cyber threat modelling process and reviewed a number of widely accepted threat modelling frameworks. With the help of this survey article, we familiarised ourselves with the concepts and terminology required to carry out the process of threat modelling. We analysed different threat modelling frameworks, methodologies and populated threat models to get knowledge about the commonly identified threats already familiar. This gave us the comprehensive understanding of the concepts of the threat modelling process. Consequently, it was deduced that threat modelling activity is highly dependent on the context, comprising of scope, approach and purpose of the threat modelling. Following this, we were able to answer our research question *SQ3* in 3 steps.

Firstly, we defined the context (scope, approach and purpose) of our threat modelling activity for the *Pre-Data Marketplace Platform 1.0* from Chapter 2. The context was defined, "*to establish the assets associated with the business functions of each functional component of the high-level architecture of a technological entity and later, assume a system specification on which applicable cyberattack vectors (described at a high-level) can be identified"*. After defining the context, we deduced that none of the frameworks or methodologies in the literature apply to our context. Evidently, we found a gap in the threat modelling literature with respect to threat modelling of the entities (information systems or organizations) at the scope of business functions.

Following this, we moved on to the step 2. We identified NGCI Apex Classification of Cyber Threat Models in which we could map our context to one of the threat models in the classification, *High-Level Threat Model.* A high-level threat model generally consists of the threats described in general terms which support the high-level risk assessment. They termed these threats as *threats of concern which* are generic threat events, narrative threat scenarios and adversary characteristics driven by assets which described briefly; which essentially relates to our defined context.

Then in step 3, Inspired from the conceptualisation of *high-level threat models* from NGCI Apex Classification, we rephrased our context into "***performing high-level threat modelling of the high-level architectures of the technological entity***" and furthermore, we resorted to *develop a framework* to obtain a high-level threat model as applicable to the context. Consequently, we designed a simple framework to carry out high-level threat modelling of high-level architectures of technological entities and hence, named it as ***High-Level Threat Modelling (HLTM) Framework.*** The framework consisted of the constructs: *Functional Component, Business Function, IT System Asset (Primary and Supporting Assets), Cyber Threat, CIA Violated, Business Consequence* and *Mitigation Technique.* It was conceptualised that the constructs, *threat* and *business consequence* reflect the ***high-level threat landscape*** of the focal technological entity. Consequently, the framework was chosen to carry out the threat modelling of the *Pre-Data Marketplace Platform 2.0,* thus answering *SQ3.* Furthermore, the framework contributes towards filling the gap identified earlier in the literature with respect to the threat modelling at the scope of business functions.

4

A New Threat Model for Data Marketplace Platforms

The second half of the research task, *RT2: To identify the threats associated with the architecture from RT1* (follow-up of Chapter 3), is discussed in this chapter; and the following research question is answered,

SQ4: What are the threats associated with the data marketplace platform from SQ2?

With the help of a literature analysis of cyberattack vectors and their consequences, the HLTM framework from Chapter 3 was applied on the *Pre-MPC Data Marketplace Platform 1.0* from Chapter 2, and a *new threat model* comprising of high-level threats to the data marketplace platform was developed. The resulting model answers *SQ4* and provided a baseline overview of threat landscape of the data marketplace platforms. The rest of the chapter discusses the high-level threat model which signifies the *Pre-MPC Threat Model 1.0*

4.1 High-Level Threat Model for the Data Marketplace Platform

In the larger research gap of the realisation of the data marketplaces, one of the gaps is with respect to the threats faced by them. Researchers have discussed the legal and economic challenges of setting up a data marketplace (Koutroumpis et al., 2017). But the threat landscape of the data marketplaces has never been explored although it represents a significant element (security aspect) towards their realisation. The researchers have touched upon such security aspect by just suggesting that the confidentiality and privacy of the data being transacted are the concerns to be explored. We went beyond this and built a comprehensive threat model comprising of all sorts of threats applicable to the high-level architecture of the data marketplace platform, thus providing a high-level overview of the threat landscape of the data marketplace platforms.

The application of HLTM framework is straightforward as discussed in Chapter 3. Firstly, each functional component was mapped to its business functions. Then, the basic IT system assets which enable the business functions are assumed based on our experience with computer science and engineering background. IT system assets are assumed according to the template of the

primary and supporting assets. The cyberattack vectors which could affect the identified IT assets are identified based on literature analysis and web search. Subsequently, the computer security property (CIA) violated and then, the consequence of the cyberattack to the focal business function or the whole entity are deduced, Finally, the appropriate mitigation technique is proposed for each cyberattack based on the literature analysis and web search. The resulting threat model represents the *Pre-MPC Threat Model 1.0* and the values of the *threat* and *business consequence* reflects the high-level overview of the threat landscape of the data marketplace platforms.

The threat model is discussed in the following subsections which are divided according to the functional components of the high-level architecture.

4.1.1 Threats: Identity Management

As discussed in Chapter 2, the main objective of the component, *Identity Management*, is to enforce the *boundary conditions*. This involves establishing strict processes to induct the customers to use the platform services and restrict access to unauthorised entities. Consequently, we established that this component involves the following business functions: *induction, authentication and authorization*. Each business function is dealt, and the corresponding threats and business consequence are discussed in the rest of the subsection.

Induction function is responsible for carrying out the screening process of the potential customers. The goal of this business function is to allow only legitimate customers to sign up for the services of the data marketplace platform. Since it is a B2B entity, the screening process should focus on establishing the legitimacy of the organization willing to sign up. The basic specification of *induction* could be that the customer must fill in the profile information on a web form which is submitted on the website. Further, the organization legitimacy could be validated by verifying its legal status with the national commercial registry database. After the verification, the customer could be provided with the access to the platform services with credentials. The primary assets here could be the customer organization's profile information and legitimacy verification service. The supporting assets enabling the business process could be the web form on a website, the communication channel and the identity database. Based on these assumptions of the IT asset specification, the threats were identified as listed in Table 8.

***Table 8**: Threats: Induction of Customers*

Primary Asset	Supporting Asset	Threat	CIA violated?	Business Consequence	Mitigation Technique
Customer Organization's Profile Information	Web Form on the website	• Identity Spoofing • Masquerading	Confidentiality of the DMP services	Induction of malicious entities as customers	2-step verification of authenticity
	Identity Database	• Database Injection Attack; • Malware	CIA of the customer identity information	Compromise of authentication service through disclosure of credentials and the services of the DMP to the attacker	• Usage of secure stored procedures over direct querying; • Anti-Malware

Customer Validation Service	Communication Channel	Eavesdropping/ Traffic Analysis	Confidentiality of profile information	Disclosure of the sensitive customer profile information	Encryption
	Verification of the website of the customer organization	Counterfeit website by attacker pretending to be a customer	Integrity of the verification service	Induction of malicious entity as the customer	Verification of certificates of the consumer organization website

Coming to the authentication function, the assets relevant here could be customer credentials and the authentication service. These could be supported by the website of the data marketplace. The threats relevant in this area are password attacks and denial of service attacks. These threats could be overcome respectively by imposing a strong password policy, and system monitoring to differentiate illegitimate requests, say from botnet, followed by tagging and isolating the source of illegitimate requests. These are listed in Table 9.

***Table 9**: Threats: Authentication*

Primary Asset	*Supporting Asset*	*Threat*	*CIA violated?*	*Business Consequence*	*Mitigation Technique*
Customer Credentials & Authentication service	Website	Password Attacks • Brute Force Attack • Dictionary Attack • Cookie Replay Attack	Confidentiality of the DMP services	Access of the DMP services to malicious entities	• Strong Password Policy • Cookie Management
		Denial of Service Attack (DoS, DDoS, Botnet)	Availability of the DMP	Inability for legitimate customers to access DMP	System Monitoring for illegitimate requests

Authorisation involves providing appropriate privileges to the applicable customers. This includes differentiating the customers and enforcing boundaries between the customers who have access to the platform services and the ones who have access to the data products that they have bought. The data products that are bought could be a one-time supply or a periodic supply or a real time continuous one. Depending on these parameters, the privileges should be managed and maintained. Configuration errors here might result in access to unauthorised entities. This can be overcome by periodic review of privileges and access controls. These access controls and privileges could also be target for external attackers to gain access to the system. These could be combatted with firewall and intrusion prevention system. This discussion is listed in Table 10.

***Table 10**: Threats: Authorisation*

Primary Asset	Supporting Asset	Threat	CIA violated?	Business Consequence	Mitigation Technique
Customer Privileges	Authorisation systems	Configuration errors caused by human errors	Confidentiality of the DMP services to unauthorised entity	Access of the DMP services to malicious entities	Periodic review of access controls and privileges
		Manipulation of privileges by attacker after entering the system	Integrity of authorization system	Privilege allocation and access controls to malicious attacker	• Firewall • Intrusion prevention system

4.1.2 Threats: Broker Service

The broker service component aims to provide the platform services to the customers through its 2 business functions: *Data Management* and *User Interaction*.

Data management service takes care of the background processes responsible for providing the data marketplace platform services: Data Cataloguing, Data Marketplace Curation and Data Tracking. These services could be carried out on a server which is supposed to be up and running 24/7. The threats applicable in this scenario could be that if the integrity and availability of the services is disrupted which could sabotage the broker operations. One of the attack vectors capable of causing this is malware. Malware attacks comprising of Viruses, worms, payloads with malicious code can manifest into processes which could disrupt the backend services potentially sabotaging the platform. This could be combatted with an updated anti-malware installed in the system along with firewall and intrusion prevention system. In addition to this, resource enumeration & browsing attack could cause damage to data management activities by disclosing the inner mechanism of the data management services to the attacker. With this attack, the attacker can learn about the resources and their configuration to plan a follow-up sophisticated attack to the systems. This could be overcome by installing a firewall with intrusion prevention system to monitor and restrict the unauthorised requests to the system. The above discussion is listed in Table 11.

***Table 11**: Threats: Backend features: Data Management*

Primary Asset	Supporting Asset	Threat	CIA violated?	Business Consequence	Mitigation Technique
Data Cataloguing, Data Curation, Data Tracking Services	Server in a data centre with the applications carrying out data management services	Malware attacks to sabotage the DMP service	CIA of the platform services	Failure of platform services	Anti-Malware with updated malware definitions
		Resource enumeration & Browsing attack	Confidentiality of Backend resources and operations	Disclosure of the backend resources to the attacker	• Firewall • Intrusion prevention system

Frontend features involve the interface services for the customers which provide them with the data marketplace experience. All the services could generally be provided through a website and the services include publish, browse, search, transform and access the (meta)data. The threats here generally could involve the ones that affect the web applications. Open Web Application Security

Project (OWASP) have researched extensively on the threat events to web applications and have published 20 threat events directed towards a number of specific web application vulnerabilities (Watson & Zaw, 2018). All the threat events mentioned in OWASP application apply here as it is web-based service but again, the threats are implementation dependent. We included a few general threats we think are crucial. Alteration attack involves tampering the source code of the website and affect its integrity to either disrupt the service or to launch a further attack. These could be restricted by safeguarding the source code from modification which links to privilege management. Further, the usual culprits affecting the CIA apply here. Denial of Service using Botnet attack vector could affect the availability of the website and frontend services to the customers. Eavesdropping/Traffic analysis and Man-in-the-Middle attacks could be used to intercept the information being transmitted in the communication channel from the website to the server or vice versa. Furthermore, the intercepting entity could alter the information to make malicious requests posing as a legitimate entity potentially disclosing sensitive information or sabotaging the Data Marketplace services. These could be overcome by encryption of the communication channel and valid certification of the website to establish the trustworthiness of the website. These threats are listed below in Table 12.

***Table 12**: Threats: Frontend features: User Interaction*

Primary Asset	Supporting Asset	Threat	CIA violated?	Business Consequence	Mitigation Technique
User Interaction Services	Website with outward facing services	Website defacement attack with alteration/ modification attacks	Integrity of the website	Faulty website with faulty functionalities resulting in reputation loss.	Restricted access to the website source code
		Denial of Service attack (DoS, DDoS, Botnet)	Availability of the website to the customers	Disruption of the website service to the customers	System Monitoring for illegitimate requests
	Communication Channel	Eavesdropping/ Traffic Analysis	Confidentiality of transmitted information	Disclosure of sensitive information	Encryption
		Man-in-the-Middle Attack	Integrity of the information and the service	• Manipulation of the sensitive information • Disclosure of sensitive information to malicious attackers posing as legitimate customers	• Encryption • Firewall • Intrusion Prevention System

4.1.3 Threats: Clearing House

The IT asset involved in this component is transaction management service which stores all the information of all the transactions happening on the data marketplace platform. This could basically be powered by a database management system and hence, the threats that apply here are database threats. These are listed in according to their applicability with the transaction management in Table 13. The compromise of transaction management service could impact the data marketplace operations to a great extent as transaction management is responsible for the

core business of the data marketplace. A compromise might lead to loss of transaction information potentially losing the track of data product being transacted. This could potentially make the platform lose the legal tracking of the product thus leading to regulatory complications. The transaction management could be safeguarded with anti-malware, firewall and intrusion prevention system to prevent external attacks while carrying out periodic maintenance and database auditing to monitor its functioning.

***Table 13**: Threats: Clearing House*

Primary Asset	Supporting Asset	Threat	CIA violated?	Business Consequence	Mitigation Technique
Transaction Management Service	Database Management	Injection Attack	CI of the transaction data	• Loss of data provenance • disclosure of customer profile information with transaction details	Usage of secure stored procedures over direct querying
		Malware	CIA of the transaction data	Disruption of the website service to the customers	Anti-Malware
		Update Errors, Incomplete transactions	Integrity of the transaction data	Loss of *data provenance* losing legal connection with the data product	Frequent Auditing of database processes

4.1.4 Threats: Data Inventory

Data inventory is the storage component of the data marketplace platform which manages and maintains the data products being transacted on the platform. Based on the design of the marketplace (*centralised and decentralised*), the data inventory differs with its implementation. Threats to both the designs are listed in Table 14.

In a centralised design, the data providers publish their data assets on to the platform transferring the data sovereignty over to the data marketplace. The data is stored by the platform and is transferred to the data consumer when the data is purchased. This involves the requirement of infrastructure for the storage of large volumes of data (Big data). Though it is implementation dependent, the big data storage is carried out with the help of data stores powered by flash storage supported by big data tools like Hadoop, Cassandra, NoSQL et cetera. These data stores are prone to threats because of the valuable commercial data they house. Because it is assumed to be a data store, the threat could not be one specific attack, rather could be mentioned as hacking comprising all the 7 steps of a generic cyberattack: Reconnaissance, Scanning, Access & Escalation, Exfiltration, Sustainment, Assault and Obfuscation (D. A. Smith, 2017). A successful attack at different stages of hacking causes damage to the data store. With respect to the assets they house i.e. data, a data breach causing the disclosure of proprietary data products published by providers on the platform could cause fatal damage to the data marketplaces in the form of financial, reputational and customer losses. If the data involved consists of the personal data collected from the users of the services provided by the data providers, the data breach can cause the violation of soft privacy leading to regulatory impacts on the data marketplace. Soft privacy refers to the violation of the privacy by an entity whose holds the personal data which is bought from other companies who directly collect from the users. The security techniques to safeguard the data on the data store could involve storing the data in the encrypted form. Furthermore, the servers need to be secured

with firewall, anti-malware, intrusion prevention systems and system monitoring which form the basic infrastructure for security in organizations.

In a decentralised design, the metadata repository is the main asset managed by the data marketplace as part of the data inventory component. The reason being the data which is sold over the data marketplace are managed and maintained by the data providers themselves and provide only metadata information of the data sets to the marketplace. The metadata information is managed by the data marketplace and uses it in its broker service to connect the supply and demand. Further, a communication channel could be set up between the transacting parties to transfer the data being purchased on the marketplace. This aspect is part of data exchange service which will be dealt in the next subsection. The metadata management could involve database management and applications run on the server as supporting assets which could be subjected to attacks like Injection or malware to disrupt the metadata management. this could cause the disclosure of metadata information result in the loss of proprietary information. With this, the customer might lose the valuable resource and could hold data marketplace legally liable. The injection attacks could be overcome by using stored procedures over letting the customers query the metadata. These threats could also apply to centralised design as it also deals with metadata management along with data storage.

***Table 14**: Threats: Data Inventory*

Primary Asset	Supporting Asset	Threat	CIA violated?	Business Consequence	Mitigation Technique
***Centralised Design**: Data assets published by the data providers; 2 variants: proprietary data and metadata*	Data Store with flash storage coupled with servers powered by Hadoop, Cassandra, NoSQL et cetera.	Hacking • Reconnaissance • Scanning • Access & Escalation • Exfiltration • Sustainment • Assault • Obfuscation	• CIA of the data sets • Integrity of the DMP service	• Data Breach causing the disclosure of proprietary data of providers to attackers causing financial, regulatory and reputational losses • Soft Privacy violation in case of private data.	• Encryption • Firewall • Anti-Malware • Intrusion Prevention System System Monitoring
***Decentralised Design**: Metadata repository of the data products, metadata contains terms of usage*	Database Management of metadata information	• Injection Attacks • Malware	• CIA of metadata • Integrity of the DMP service	• Disruption of the metadata management • Disclosure of metadata information of datasets of customers revealing metadata information which can be proprietary, contractual information etc.	• Stored Procedures • Encryption • Anti-Malware

4.1.5 Threats: Data Exchange Service

This component merely signifies the transfer of the data from the data provider to the data consumer. The 2 designs (centralised and decentralised) apply here too. But in either of the designs, the threats remain the same as the core operation is the same: the transfer of large volumes of data through communication channel. In a centralised design, the communication channel between the data provider and the data marketplace; and between the data marketplace and the data consumer is the supporting asset. In the case of decentralised design, the communication channel set up between the data actors after they are matched on the data marketplace platform is the supporting asset. The threats to this supporting asset could involve the generic threats to the communication channel like eavesdropping, man-in-the-middle attacks as described in Table 15. A compromise in this area is very fatal for the data marketplaces as large volumes of commercial proprietary data are being transferred in this component. A data breach here could have the same impact as we discussed in the previous component resulting in violations of privacy agreements, loss of business-specific confidential data and so on. These threats could be mitigated by adopting a more sophisticated and secure mechanism to transfer the data between the parties. Common encryption methods could also pose risk since the resource involved is a significant one. More than just encryption, the business process of how the data assets are handled could be designed in a secure way with sophisticated security technologies.

***Table 15**:Threats: Data Exchange Service*

Primary Asset	*Supporting Asset*	*Threat*	*CIA violated?*	*Business Consequence*	*Mitigation Technique*
• ***Data being transacted*** • ***Data transfer mechanism***	Communication channel	• Eavesdropping / Traffic Analysis • Man-in-the-Middle • Malware	Confidentiality of the data; Integrity of the transfer service.	• Data Breach causing the disclosure of proprietary data of providers to attackers causing financial, regulatory and reputational losses. • Soft Privacy violation in case of private data.	Encryption

4.1.6 Threats: Data Analysis Service

The business function assumed for this component in our architecture is like that of an app store. Here, in addition to the data marketplace providing its own data analytics tools. It could allow third parties to upload their big data analytics tools and offer them to the customers of the data marketplaces. In this setting, the threats we could think of are with respect to the authenticity of the third-party data analytics tools. The tools could be uploaded by malicious third parties and hence, the tools can contain malicious constituents. This could cause damage to the data sets subjected to analysis by the said tools resulting in a damage to the customer and in turn to the data marketplace in terms of legal liability and reputational deterioration. We could mimic an actual app store approach to overcome this threat by incorporating quality and security checks to the tools being provisioned by third parties. This way the customers could judge the authenticity of the services and trust the data marketplace. The above discussion is represented in Table 16.

***Table 16**: Threats: Data Analysis Service*

Primary Asset	Supporting Asset	Threat	CIA violated?	Business Consequence	Mitigation Technique
***Data Analytic Tools**: either downloadable or provided as SAAS*	Third party analytics tools uploaded on marketplace similar to app store.	• Faulty Software • Malicious software uploaded by a malicious third party.	Integrity of the app store service of the data marketplace	Reputation loss and Legal liability for providing customers with malicious or faulty analytics tools	Screening and quality check of the analytics tools published by the third parties.

This marks the end of the threat modelling activity. The resulting *high-level threat model* which is represented by the tables: Table 8, Table 9, Table 10, table 11, Table 12, Table 13, Table 14, Table 15 and Table 16; collectively answer the sub-research question, *SQ4*. The threat model satisfied our context in the sense that the threats are described at high-level to the high-level business functions of the data marketplace platform. Furthermore, the combination of the threats and business consequences of all the business functions associated with every functional component of the *Pre-MPC Data Marketplace Platform 1.0* represents the high-level threat landscape of the data marketplace platform. Furthermore, the high-level threat model signifies the *Pre-MPC Threat Model 1.0*

Although a pioneer effort towards exploring the threat landscape of the data marketplace platforms, because of the limitation of the *HLTM framework*, our work in this chapter only represents a ***baseline overview of the threat landscape of the data marketplace platform***. For this, reason, the threat model was subjected to validation later in Chapter 7 to make it further from being just a baseline overview and to obtain a more valid overview representing the actual threat landscape of the data marketplace platforms.

5

Effect of MPC on Architecture and Threat Landscape of Data Marketplaces

This chapter marks the end of the *Conceptualisation phase* which involves the research task, *RT3: To investigate the effect of MPC technology on the architecture from RT1 and the threat model from RT2*; and the following 3 sub-research questions are answered.

SQ5: *How to incorporate MPC technology into the architecture of the data marketplace platform from SQ2?*
SQ6: *What is the effect of MPC incorporation on the rest of the architecture from SQ2?*
and
SQ7: *What is the effect of MPC incorporation on the threats associated with the data marketplace platform from SQ4?*

SafeDEED project proposal is studied to understand its plan to promote the data sharing culture among the organisations with the incorporation of their technologies and specifically focussed on their conceptualisation of MPC technology and how they intend to materialise its processes. These processes were crucial to understand for it to be incorporated into the data marketplace platform, thus answering *SQ5*. The proposed processes were attempted to position in the literature but were not specifically found. The processes were incorporated anyway into the high-level architecture from Chapter 2 and its effect on the rest of the architecture was analysed to obtain an updated architecture with reflecting MPC incorporation, which answers *SQ6* and signifies *Post-MPC Data Marketplace Platform 1.0.* Following this, the effect of MPC incorporation on the threat model from Chapter 4 was deduced by analysing the implication of the MPC incorporation on each threat in the model related to the functional components of the architecture. This answers *SQ7* and the resulting threat model signifies the *Post-MPC Threat Model 1.0.*

The rest of the chapter is structured as follows. Section 5.1 describes the SafeDEED's agenda to enable safe and secure inter-organizational data sharing. Section 5.2 discusses the concept of MPC

technology and SafeDEED's proposed implementation of MPC technology along with its 2 variants of processes. Section 5.3 discussed the incorporation of MPC technology and its effect on the *Pre-MPC Data Marketplace Platform 1.0.* Section 5.4 discusses the effect of MPC incorporation on the *Pre-MPC Threat Model 1.0.* Section 5.5 summarises the chapter where the focal sub-research questions, *SQ5, SQ6* and *SQ7* are formally answered.

5.1 SafeDEED: Safe Data Enabled Economic Development

The plan of SafeDEED project to boost the data market (not the data marketplaces) to foster the data economy in Europe is discussed in this section. The corresponding information is adopted from the project proposal of SafeDEED**,** written by Mihai Lupu (2018).

SafeDEED aims to develop a set of technologies to incentivise and enable data providers into sharing their data to other companies in need, thus creating value for both the sides of the data market; ultimately, fostering data-driven business model innovation. SafeDEED believes that the data market of Europe has the potential to help organizations to keep up with international competition but suggests that the sharing of data among organizations is hampered by the following barriers (Lupu, 2018):

- Lack of trust in data suppliers and data aggregators
- Lack of awareness of data sharing and business opportunities.
- Organizations fear to lose power/control of owned data.
- Enterprises' uncertainty in the implementation of GDPR.

These barriers are the ramification of the sensitivity existing around the tricky nature of commodification of data which includes privacy concerns (in case of private data), confidentiality breach (in case of commercial proprietary data), intellectual property enforcement challenges et cetera. SafeDEED aims to overcome these barriers by developing 2 categories of technologies: Multi-Party Computation (MPC) and Data Valuation Technologies (DVT). With these technologies, SafeDEED aims to solve the problem of reluctance that exists in data owners with respect to sharing data to external entities. MPC category is a component which enables the data owners to share their data in a confidentiality-preserving and privacy-preserving manner. This is the focal technology of this chapter and also this thesis. However, it is not analysed only for its privacy preserving nature but for the comprehensive promise it brings as a technology to the architecture from Chapter 2 and generally, to the species of data marketplaces. This will be discussed further in the rest of the chapter. On the other hand, DVT are the technologies which explicate the value of the data and thereby, encourage the data owners to commoditise their data to encash on that value; while making the data appealing to the data consumers. The latter category of technologies is not relevant to our scope and hence, is not dealt here.

5.2 MPC Technology & *SafeDEED Component*

Secure computation is the solution for the famous problem called "*Two Millionaires problem*" where 2 millionaires wish to know who is richer without disclosing information about each other's wealth. Yao (1982) designed a protocol which solves this problem and it does so with the help of secure computation. The same solution has been researched to include more parties such that multiple parties can compute functions on the union of their data to produce desirable output without having to merge the individuals' actual data (Goldreich, 1998). This functionality finds an application in the context of data market where data security is a crucial aspect.

Multi-Party Computation (MPC) is a type of cryptographic protocols which allow functions to be computed over distinct datasets without having to share the data itself. As a result, the required knowledge from the data can be extracted without revealing the actual data. This characteristic is appealing to the data owners to create value as with MPC, they can share the business intelligence of their data without giving access to actual data. Several MPC methods already exist which carry out the above-mentioned functionality with mathematical sophistication. However, they suffer from scalability and performance limitations which restrict their usage in real-world applications. SafeDEED claims to overcome these limitations and provided a practical solution which will be testied with pilot cases. SafeDEED claims to develop faster MPC protocols viable also for larger data sets by improving the computational and communication complexity of the underlying technical components.

To perform computation on the datasets using MPC protocols, it is necessary to know the function beforehand that needs to be applied on the data. The function signifies the knowledge that needs to be extracted from the data. Based on this function, the corresponding MPC protocol which can perform this function can be designed by selecting appropriate technical components. For example, if multiple companies want to perform mean and variance on some of their customer's data, then the functions, mean and variance need to be represented as circuit using addition and multiplication gates. These addition and multiplication gates constitute the technical component blocks for building the MPC protocol. To help this cause, SafeDEED proposes to develop those technical components required to execute different protocol of different functions. These technical components are referred as ***SafeDEED Primitives***. These consist of convenient and easy-to-use methods to build protocols for the required function without requiring the deep understanding of the underlying technical aspects. These primitives involve cryptographic building blocks like low multiplicative complexity symmetric-key, garbled circuits, oblivious transfer and so on; which will be selected according to the requirements in designing the protocol. The designed protocols need to support communication and hence, SafeDEED also provides a network component powered by transportation libraries such as OpenSSL or GnuTLS, which they refer as ***SafeDEED Network.*** The whole offering of SafeDEED comprising of the constituents, *SafeDEED Primitives* and *SafeDEED Network* is referred as ***SafeDEED Component*** (Lupu, 2018) and is as illustrated in the schematic diagram in Figure 11.

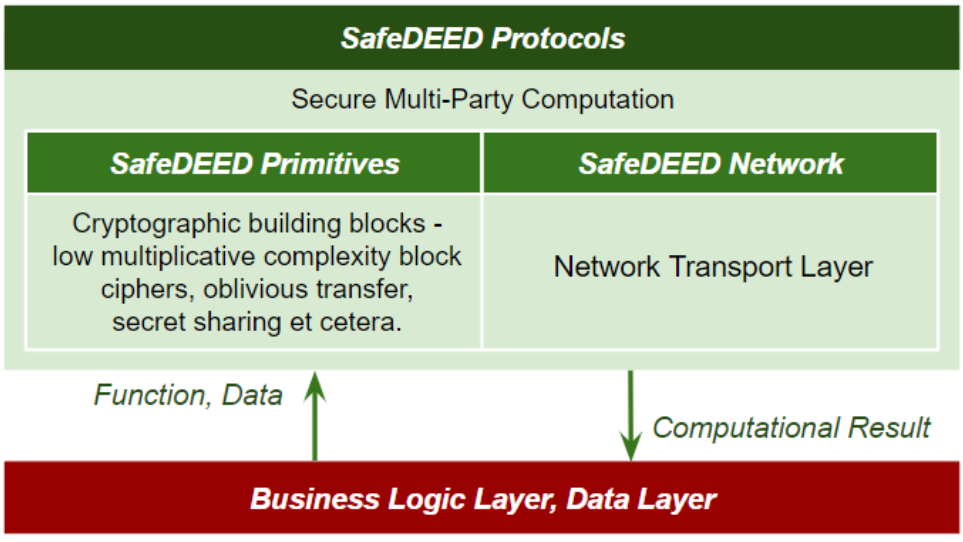


***Figure 11**: **SafeDEED Component** for MPC Technology; Source:* Lupu (2018)

SafeDEED Component acts as a black box accepting the specification of the function and the data; and generates computational result which reflects the required outcome expected from the union

of the data of the parties involved Basically, SafeDEED wishes to simplify the design of MPC protocols where the user who adopts the *SafeDEED Component* only needs to decide on the function to be evaluated with other parties and has to supply the input data. Further, SafeDEED takes care of the underlying technology in designing the protocol with the appropriate technical blocks.

5.2.1 MPC processes proposed by SafeDEED

The concept of MPC protocols discussed for are interactive approaches where the parties involved should have their data available simultaneously with all the actors for the computation to happen; i.e. in a synchronous way. This kind of process is represented in the schematic diagram in Figure 12.

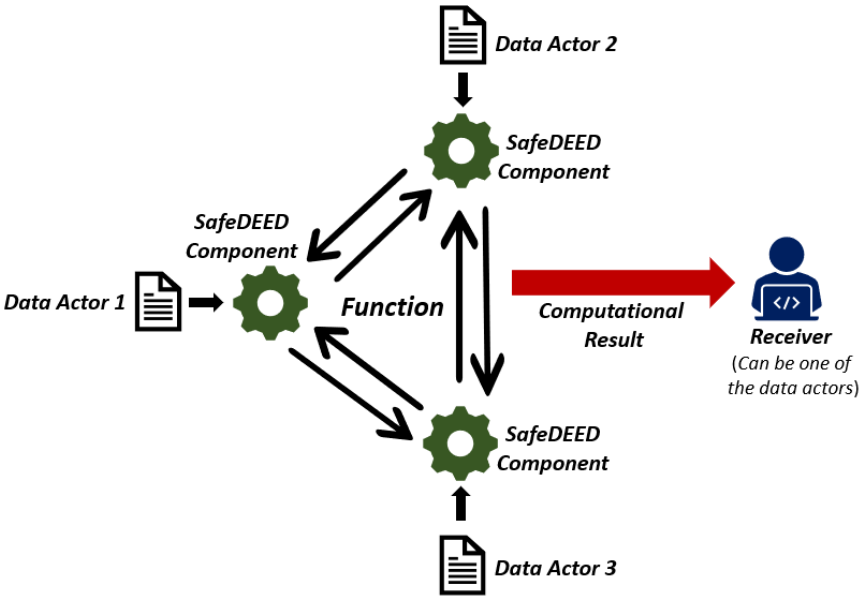


***Figure 12**: Interactive MPC Process*

SafeDEED also explores non-interactive approaches where the data sharing can happen in an asynchronous way. SafeDEED proposes homomorphic encryption as the enabler of this kind of asynchronous data sharing. Homomorphic encryption is a variant of encryption scheme that allows one to evaluate functions on encrypted data. SafeDEED proposes a case where data providers encrypt their data to a dedicated receiver and send it to a dedicated aggregator who then evaluates the function on the ciphertexts and forwards the computational result to the dedicated receiver. This kind of process is referred to as *multi-user data aggregation scheme* and this is illustrated in Figure 13. In this way, the process provides a non-interactive approach for data sharing which enables the providers to share data in an encrypted form which can be used later by the dedicated actors without having demand the presence of the data provider.

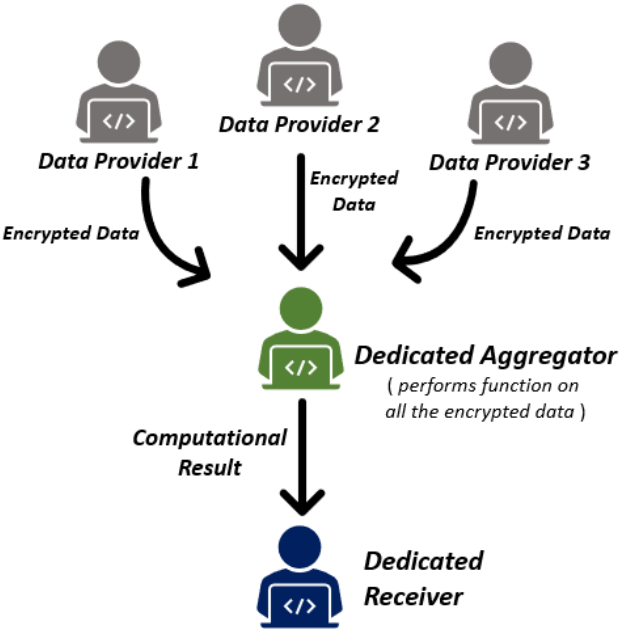


***Figure 13**: Non-Interactive MPC Process*

These 2 approaches are supported by the literature dealing with the application of MPC and Homomorphic Encryption in the data marketplaces. Roman & Stefano (2016) designed a concept, *Trusted Data Marketplace* operating solely for the application of credit scoring. They design a reference architecture for a data marketplace platform where the actors involved in credit scoring can trade their data among each other. They suggest homomorphic encryption and multi-party computation as enabling technologies for the realisation of their concept data marketplace where the physical data either remains with the data owner or is in encrypted form (by Homomorphic Encryption) stored on a cloud. They discuss 2 settings of data mining powered by MPC.

- In the first scenario, the data is held by 2 or more different parties and the data mining algorithm is run on the union of these parties' databases without letting each other know of other's data. This setting reflects the traditional MPC process where a function is computed on the union of databases from multiple parties to get a result.
- In the second scenario, some statistical data needs to be released for research or data mining. But the data might contain private information, hence, the data is modified first perhaps with anonymization so that the privacy is not compromised and parallelly meaningful results can be obtained from the anonymised data. This is a special case of the first scenario where, the parties anonymise their data before lending it for the computation where the MPC protocol carries out the union and the function execution.

These 2 scenarios reflect only the first out of the 2 processes suggested by SafeDEED. Since, the work of Roman & Stefano (2016) is the only article we could find as of the date ***14 June 2019*** which deals with the application of MPC technology in data marketplaces, we cannot say for sure that the processes apply in data marketplaces.

However, both the processes can be implemented within the *SafeDEED Component* and this could be integrated as a component or a feature into the architecture. In this way, *SafeDEED Component* provides a way of incorporating the MPC technology into the high-level architecture from Chapter 2; thus, answering *SQ5*. The logistics of integration is discussed in the next section.

5.3 MPC Incorporation into the Data Marketplace Platform

Here, the SafeDEED component is integrated into the high-level architecture of the data marketplace platform from Chapter 2; furthermore, the effect of its integration is also analysed.

The concept of MPC protocol could be related as a mechanism of the transferring the knowledge within the data from the data provider to the data consumer (without transferring the actual physical data). Consequently, the SafeDEED component can be viewed as a component which enables the process of data exchange and hence, SafeDEED Component was integrated into the ***Data Exchange Service*** of the high-level architecture as its business process.

The incorporation essentially makes the data marketplace platform a purely decentralised one as no physical data transfer is involved. Essentially, the platform will be responsible just for connecting the data providers, data aggregators and data consumers Following the establishment of the relationship between the actors over the platform and the *Data Exchange Service* powered by *SafeDEED Component* would be set up by the marketplace ad-hoc between the dedicated data actors outside the platform for them to interact with each other and share data. Furthermore, the computation of the function on the data from the involved actors will be performed by the *SafeDEED Primitives (MPC Protocol)* according to the requirement. The computational result is then presented to the dedicated receiver through the communicational channel powered by *SafeDEED Network*.

Furthermore, there would be no need for a ***Data Inventory*** within in the architecture as the platform is decentralised now. So, the component gets transformed into just *Metadata Inventory* which just stores the metadata of the data provisioned to be transacted over the platform and will be used by the *Broker Service* which showcases the metadata to the customers through its functions. The backend features of *Broker Service* component also go through changes where the management activities like cataloguing and curation activities are done only for the metadata of the data. Since there will be no data publishing on the platform, the data aggregator steps out of the umbrella of data providers. The aggregator's function with respect to this design is aggregation of the data and not publishing the aggregated data. Hence, the data aggregator becomes a distinct actor who will avail the platform to provide his aggregating services. Meanwhile, the data provider actor is transformed into just data owner who holds the different types of data like raw data, polished data, formatted data et cetera and provisions the data on the platform by publishing its metadata. The actors, *Data Collector* and *Data Manager* considered earlier now fall under *Data Owner* as they own and offer data on the platform.

With respect to the functional requirements, *Secure Data Exchange* requirement is enabled by *SafeDEED component* with its MPC protocols. *Data Sovereignty* is retained by the data provider as the provider holds the control over his physical data. *Data Governance* is also taken care of by the data provider as he becomes responsible for the management and maintenance of his data. The modified high-level architecture of the data marketplace platform after the incorporation of the *MPC technology,* signifying the *Post-MPC Data Marketplace Platform 1.0* is illustrated in Figure 14. with the modified elements highlighted in ***yellow***. The *Data Exchange Service* is depicted separately in Figure 15 which reflects the functioning of *SafeDEED Component.* This updated architecture answers *SQ6*.

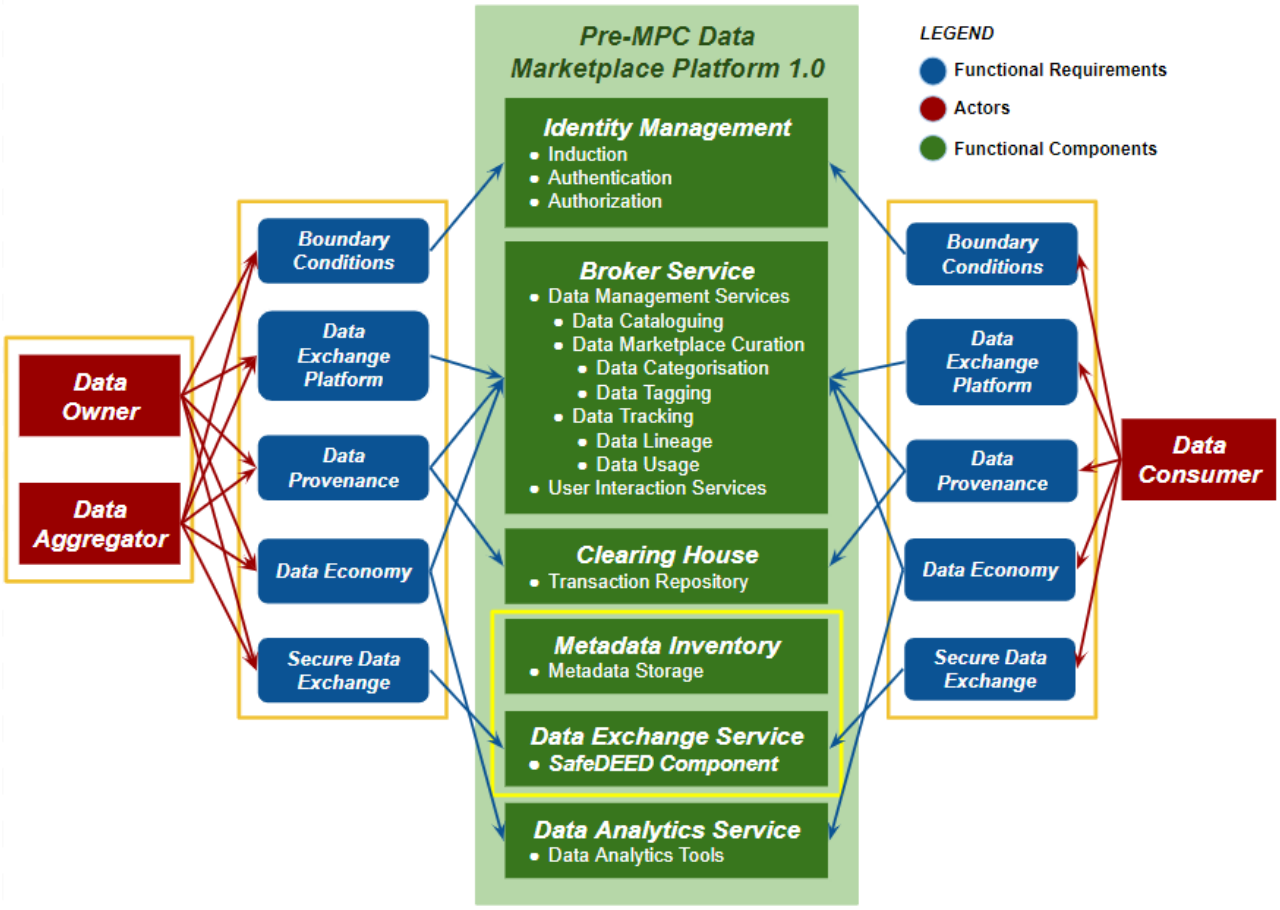


***Figure 14**: Post-MPC Data Marketplace Platform 1.0*

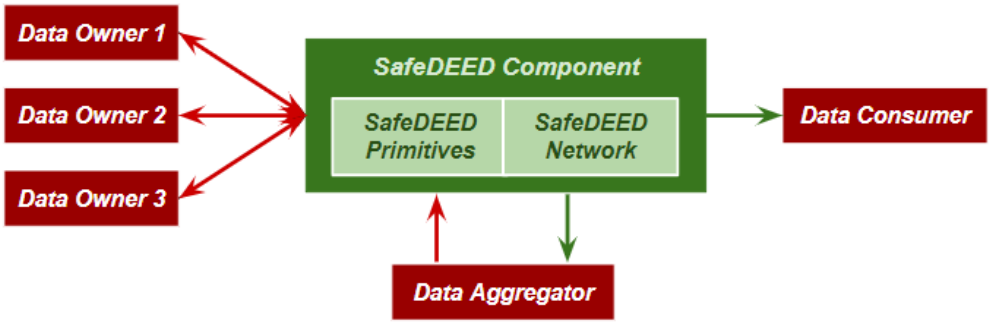


***Figure 15**: Data Exchange Service enabled by SafeDEED Component powered by MPC*

5.4 Effect of MPC Incorporation on the Threat Model

The functional components that undergo major change with the incorporation of MPC technology are Data Inventory, which is now, *Metadata Inventory* and *Data Exchange Service.* As a result, the incorporation affects the threats of only these 2 components and not that of any other component in the architecture.

5.4.1 Post-MPC Threats: Metadata Inventory

Since the platform now is decentralised, the commercial proprietary data stays at the site of the data owner and there is no transfer of physical data over the platform, the incorporation overcomes the risk of data breach or the violation of privacy (in case of private data). The platform now houses only the metadata of the data provisioned by data owners. The threats identified in the threat model still apply to this metadata. However, the risk with the disclosure of the metadata is less compared to the disclosure of the commercial data. This way, the risk involved with the inventory is reduced by the incorporation of MPC technology in the data marketplace platform. The threats associated with the modified *Metadata Inventory* component is listed in the Table 17.

***Table 17**: Post-MPC Threats: Metadata Inventory*

Primary Asset	Supporting Asset	Threat	CIA violated?	Business Consequence	Mitigation Technique
Metadata repository of the data products, metadata contains terms of usage	Database Management of metadata information	• Injection Attacks • Malware	• CIA of metadata • Integrity of the DMP service	Disruption of the metadata management, disclosure of metadata information of datasets of customers revealing metadata information which can be proprietary, contractual information etc.	• Stored Procedures • Encryption • Anti-Malware

5.4.2 Post-MPC Threats: Data Exchange Service

Since the data exchange now would happen via *SafeDEED Component* in the form of an MPC Protocol, the threats causing the data breach, impose lesser risk as the information in the communication channel is either an intermediate result obtained during the protocol execution or but not the actual data. So, when the communication channel is compromised by an outsider's attack, the breached information will not be of any use to the attacker as the physical data is not there. However, the threats causing the breach of the communication channel disrupts the data marketplace service, compromising its integrity. The threats associated with the modified *Data Exchange Service* are listed in Table 18.

***Table 18**: Post-MPC Threats: Data Exchange Service*

Primary Asset	Supporting Asset	Threat	CIA violated?	Business Consequence	Mitigation Technique
• ***Data being transacted*** • ***Data transfer mechanism***	Communication channel powered by ***SafeDEED Component***	• Eavesdropping/ Traffic Analysis • Man-in-the-Middle • Malware	Integrity of the data transfer service and in turn, integrity of the DMP service.	Service disruption of DMP	Intrusion Prevention system.

Apart from these 2 components, according to our analysis MPC technology do not address the threats in the rest of the components. They still prevail and the proposed mitigation techniques must be adopted for those threats. As a result, Table 15 and table 16 along with the threat models of the rest of the components, Table 8, Table 9, Table 10, Table 11, Table 12, Table 13 and Table 16; represent the ***Post-MPC Threat Model 1.0***. Ultimately, it can be stated that MPC technology increases the security value of the data marketplace platform by addressing the most significant

factors associated with data handling on the data marketplace platform in a *Security-by-Design* way; which is the answer for SQ7.

5.5 Summary

In this chapter, we discussed the research task, *RT3* whose purpose was to understand the effect of the incorporation of MPC technology on the architecture and the threats associated with the data marketplace platform; which was accomplishes by answering the following sub-research questions,

SQ5: *How to incorporate MPC technology into the architecture of the data marketplace platform from SQ2?*
SQ6: *What is the effect of MPC incorporation on the rest of the architecture from SQ2?*
and
SQ7: *What is the effect of MPC incorporation on the threats associated with the data marketplace platform from SQ4?*

SafeDEED project proposal was studied to understand SafeDEED's plan to promote the data sharing culture among the organisations with the incorporation of their technologies; specifically focussing on their conceptualisation of MPC technology and how they intend to materialise its process. It was deduced that SafeDEED materialises MPC technology with its ***SafeDEED Component*** comprising of the ***SafeDEED Primitives,*** which provides the technical blocks required for building the protocol and ***SafeDEED Network,*** which provides a communication channel for the execution of the protocol. This *SafeDEED Component* provides a black box way of incorporating MPC technology for the customers who could just choose the required function and provision and let the SafeDEED Component to build and execute the protocol. Hence, SafeDEED Component answers *SQ5.*

SafeDEED Component was integrated into the *Data Exchange Service* as they both represented a mechanism of transferring data or the knowledge inside it from the data owner to the data consumer. As a result, the platform would become decentralised where the actors can meet over the platform and the *Data Exchange Service* enabled by *SafeDEED Component* is set up ad-hoc by the marketplace outside the platform the actors to execute the protocol and share data. This move also eliminated the need for *Data Inventory* which now is transformed into Metadata Inventory which stores and maintains metadata of the data provisioned on the platform. Furthermore, the requirements of *secure data exchange* are reinforced; while *data governance* and *data sovereignty* are moved to the site of the actor owing the decentralised transformation of the platform. Furthermore, there is a change in the way the customers are represented and now they are comprised of Data Owners, Data Aggregators and Data Consumers. This collectively is the effect of MPC incorporation into the high-architecture of the data marketplace platform and hence, the answer for SQ6. The resulting updated architecture represents the *Post-MPC Data Marketplace Platform 1.0.*

The effect of this MPC incorporation on the threat model from Chapter 4 is that this move minimises the risks associated with the components, *data inventory* and *data exchange service* as the element of physical data is eliminated from the components. Apart from these components, MPC does not interfere with the threats of rest of the components. *Ultimately,* MPC technology increases the security value of the data marketplace platform by addressing the most significant factor, data handling on the data marketplace platform in a *Security-by-Design* way; which is the answer to *SQ7*. The resulting refined threat model represents the *Post-MPC Threat Model 1.0*.

This marks the end of our *Conceptualisation phase*. The resulting artefacts from this phase are,

- HLA Framework
- High-Level Architecture of a generic Data Marketplace Platform (*Pre-MPC Data Marketplace Platform 1.0*)
- HLTM Framework
- High-level Threat Model for the data marketplace platform (*Pre-MPC Threat Model 1.0*)
- MPC Incorporated High-Level Architecture (*Post-MPC Data Marketplace Platform 1.0*)
- MPC affected High-Level Threat Model (*Post-MPC Threat Model* 1.0)

The conceptual models, *Pre-MPC Data Marketplace Platform 1.0, Pre-MPC Threat Model 1.0, Post-MPC Data Marketplace Platform 1.0 and Post-MPC Threat Model 1.0*; form the basis for the next phase of ***Validation*** where all the theoretical concepts associated with these models are validated and updated to obtain relatively more valid theoretical concepts and more valid conceptual models.

6

Validation Methodology

This chapter marks the start of the *Validation phase* where all the artefacts and their respective theoretical concepts developed during *Conceptualisation* phase are validated (*refined, updated or modified*) to generate more valid artefacts and valid concepts. With this agenda, a qualitative study was conducted by interviewing the experts in the 3 subject areas: *data marketplaces, threat modelling* and *MPC technology.* Prior to executing this study, the research task, *RT4: To design the methodology for conducting validation;* is carried out to obtain a research methodology for the qualitative study. This is dealt in this chapter where the following sub-research question is answered.

SQ7: How to validate the artefacts and their theoretical concepts obtained from the conceptualisation phase?

The research methodology was formulated by establishing its different parameters: *Design, Participants, Procedure* and *Analysis* as suggested by Kraus, Fiebig, Miruchna, Moller, & Shabtai (2015). These form the sections of the rest of the chapter; which collectively answer the sub-research question *SQ7*.

6.1 Design

The research strategy generally employed by researchers for theory development is *Grounded Theory.* Grounded Theory is a strategy to derive a theory inductively from the data (Corbin & Strauss, 1990). The process involves generating a theory by collecting the data, analysing the data which directs what data to collect next until a saturation is reached; finally, to end up with an inductively derived theory. In Grounded Theory, the theory is derived solely from the collected data. Hence, Grounded Theory can be an extreme way which truly builds a theory. However, there is a less extreme variant of Grounded Theory called, ***Middle Ground Approach*** which refines an already existing theory (Sekaran & Bougie, 2013; de Reuver, 2019). This method necessitates an initial list of codes and categories informed by an already existing theory which directs both the data collection and then, the data analysis process. This approach is a perfect fit for our research agenda of validating the artefacts from the conceptualisation phase. Hence, we adopted *Middle-Ground Approach* for our research. The first iterations of the 4 conceptual models from the conceptualisation phase constituted the *initial list of codes and categories* which also directed the design of the interview questions and thereby, the data collection. The process of the data analysis remains the same which involves constant comparison of newly collected data with the existing

list of categories and codes and then updating the theory to reflect the insights from all the collected data until theoretical saturation is reached. Our analysis is discussed in section 6.4.
The initial setup of the Middle Ground Approach i.e. explicating the initial set of categories and codes is performed first before getting into the actual methodology. It has been established in Chapter 1 that our research is related to the ***3 subject areas (SA)***:

- the new phenomenon of data marketplaces
- the threat modelling
- the new technology of Multi-Party Computation (MPC)

Related to these subject areas, ***4 research foci (RF)*** were formulated which signify the validation agendas for the *Artefacts 1.0* of the conceptualisation phase which are listed as follows,

- to validate *Pre-MPC Data Marketplace Platform 1.0* and generate *Pre-MPC Data Marketplace Platform 2.0* (*SQ8 & SQ9*)
- to validate and refine the concept of MPC Incorporation into the data marketplace platform and further, generate *Post-MPC Data Marketplace Platform 2.0* (*SQ10 & SQ11*)
- to validate *Pre-MPC Threat Model 1.0* and generate *Pre-MPC Threat Model 2.0* (*SQ12 & SQ13*)
- to deduce the effect of MPC incorporation on the threats from the *Pre-MPC Threat Model 2.0* and further, generate *Post-MPC Threat Model 2.0* (*SQ14 & SQ15*)

Relating these above-mentioned research foci, ***10 Topics (T)*** were identified which are listed as follows,

- data marketplace platform designs (*RF1*)
- functional requirements of the data marketplace platform (*RF1*)
- customers of the data marketplace platform (*RF1*)
- functional components of the data marketplace platform (*RF1*)
- HLA framework (*RF1*)
- perception of MPC technology from conceptualisation phase (*RF2*)
- MPC incorporation into the data marketplace platform (*RF2*)
- HLTM framework (*RF3*)
- threat landscape of the data marketplaces *(reflected by the threats and business consequences in Pre-MPC Threat Model 1.0)* (*RF3*)
- effect of MPC incorporation on the threat landscape of the data marketplace platform (*which is the validation of Post-MPC Threat Model 1.0*) (*RF4*)

Validation activity was performed for each one of these 10 topics using the *Middle-Ground approach*. In each topic, we dealt with several ***theoretical concepts*** which were used to answer the corresponding sub-research questions of the conceptualisation phase. These concepts comprised of *definitions, interpretations, descriptions, taxonomies, architectures, frameworks, threat models, processes* et cetera. Basically, these included every concept associated with the resulting artefacts from the conceptualisation phase. These theoretical *concepts* and their corresponding low-level information in each topic constituted the ***initial list of categories(C)*** and ***codes(C*)*** for that topic's validation activity. We derived 10 sets of initial lists of categories and codes for the 10 topics and their corresponding theoretical concepts and these 10 lists collectively represent the initial specification of categories and codes required for the Middle-Ground Approach methodology. This prerequisite information formulated prior to starting the validation phase is illustrated in the form of a hierarchy in Figure 16.

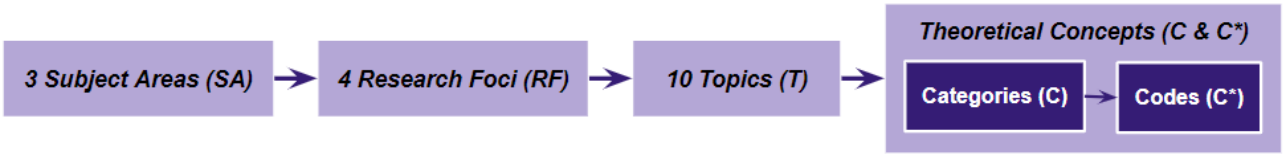


***Figure 16**: Initial Specification for Middle-Ground Approach*

Furthermore, the list of the topics associated with each research focus mapped to their respective subjective areas is shown in Table 19.

***Table 19**: Subject Areas, Research Foci and Topics*

Subject Area	Research Focus	Topic
SA1: Data Marketplaces	***RF1**: to validate Pre-MPC Data Marketplace Platform 1.0 and generate Pre-MPC Data Marketplace Platform 2.0*	***T1:** Data Marketplace Platform Designs*
		***T2:** Functional Requirements of the Data Marketplace Platform*
		***T3:** Customers of the Data Marketplace Platform*
		***T4:** Functional Components of the Data Marketplace Platform*
		***T5** HLA Framework*
SA1: Data Marketplaces & SA3: MPC Technology	***RF2**: to validate and refine the concept of MPC Incorporation into the data marketplace platform and further, generate Post-MPC Data Marketplace Platform 2.0*	***T6:** Perception of MPC Technology*
		***T7:** MPC Incorporation into the Data Marketplace Platform*
SA1: Data Marketplaces & SA2: Threat Modelling	***RF3**: to validate Pre-MPC Threat Model 1.0 and generate Pre-MPC Threat Model 2.0*	***T8:** HLTM framework*
		***T9:** Threat Landscape of the Data Marketplaces*
SA1: Data Marketplaces, SA2: Threat Modelling & SA3: MPC Technology	***RF4**: to deduce the effect of MPC incorporation on the threats from the Pre-MPC Threat Model 2.0*	***T10:** Effect of MPC Incorporation on the Threat Landscape of the Data Marketplace Platform*

The initial list of categories and codes comprising of theoretical concepts are listed in Chapter 7 when dealing with each topic so that it is easier to refer them then and there and easily compare with their updated counterparts.

6.1.1 Expert Interviews

The research method for collecting qualitative data was chosen to be ***Expert Interviews***. Interview method is one of the primary qualitative data collection methods which is widely used to collect rich data for exploratory studies in general business setting (Sekaran & Bougie, 2013). Expert interviews are a specific kind of interviews where subject area experts are specifically interviewed to obtain expert knowledge about the focal subject area. Given the research foci of our objective is related to new subject areas of which knowledge is not out there yet, we have adopted expert interviews to be our qualitative data collection method as only experts can provide insights regarding these new subject areas.

Regarding the type of interviews, it was decided to adopt ***Semi-Structured Skype Interviews***. Semi-structured interviews are the ones with some pre-defined open ended questions in an order which helps in establishing the focus on a subject while giving the flexibility to explore deeper into the subject through a follow-up discussion for the questions (de Reuver, 2019). Since our purpose of doing qualitative data analysis is to validate the concepts and artefacts from the conceptualisation phase and to update them with deeper insight, we adopted the semi-structured approach for the interview protocol. The interview questions were prepared by basing the questions on the respective concepts present in the initial list categories and codes associated with each topic's exploratory study. This way the initial list of categories and codes served their purpose in the Middle Ground approach which is to direct the data collection activity; in this case, interviews. The questions helped to explore each concept deeper while clarifying sketchy insights with follow-up questions; most of the times turning the interview into brainstorming session on the focal subject area. The interview protocol used for each expert along with the interview transcript can be found in Appendix A.

6.2 Participants

We carried out *judgement sampling* to choose the participants as it fit our objective of obtaining expert knowledge on the subject areas. Judgement sampling is a variant of purposive sampling which is used when specialized information is necessary for the study which is not available easily as that information is not mainstream (Sekaran & Bougie, 2013). The experts in the 3 subject areas: *Data Marketplaces, Threat Modelling* and *MPC technology* were considered for the interviews. The profiles of each subject area expert were formulated as follows,

- ***Data Marketplaces*:** Researchers working in the field of data intermediaries, data exchange mechanism and data marketplaces
- ***Threat Modelling***: Researchers and industry experts working in the cybersecurity domain
- ***MPC technology***: Researchers working in the SafeDEED: Safe Data Enabled Economic Development project who are conceptualising and developing the MPC technology.

The interview prospects were referred by our professors, *Mark de Reuver* and *Tobias Fiebig* who are also the in-charge of TU Delft's share of research for the SafeDEED project. The prospects were invited for the interviews with email invitations informing the experts beforehand the kind of work being dealt and what was expected of them; before they accepted the invitation. These invitations along with the interview protocols can be viewed in Appendix A. Since the purpose of the interviews was to validate the artefacts from the conceptualisation phase, it was necessary to familiarise the experts with the concepts associated with the relevant topics beforehand so that they would have better context and understanding of the concepts before getting into the interview; thereby potentially increasing the chances of their answers to be more informed and nuanced. For this purpose, the descriptions of the artefacts (as relevant for each prospective expert's subject area) consisting of the concepts were compiled into a document and was sent as an attachment with

the email invitation to the respective subject area experts. Out of 10 invited prospects, the experts who responded and were eventually interviewed are listed in Table 20 along with their relevance to our research.

***Table 20**: Experts interviewed for the Validation Phase*

Expert (E)	Role	Relevance
E1: Reggie Cushing	Post Doc researcher at University of Amsterdam. Working on a project called DL4LD (Data Logistics for Logistics Data) dealing with the conceptualising of a data marketplace in the airline industry.	Expertise in data exchange mechanisms and data marketplaces.
E2: Mihai Lupu	Research Coordinator of SafeDEED. Working closely with research partners to develop the enabling technologies for B2B data sharing like MPC, Data Valuation etc. Also working closely with Data Market Austria in its conceptualisation.	Experience in materialising a real-life data marketplace, Data Market Austria.
E3: Swati Manocha	Manager in the domain of Cybersecurity and Privacy at EY. Provides auditing and security assessment services to business clients.	Expertise in threat assessment and security frameworks.
E4: Sebastian Ramacher	Researcher in SafeDEED. Works on the implementation of Multi-Party Computation (MPC).	Expertise in MPC technology and its applications.

However, one limitation here was that 2 out of the 4 interviewees were the internal members of SafeDEED project; namely, *E2:* who dealt mainly with the area of data marketplaces and *E4:* who dealt exclusively with the subject of MPC technology. This means these both experts would provide the insights on what we already know with further clarifications; but lacking an outsider's perspective which could provide unknown but relevant insights. This limitation was overcome to some extent as we got an outsider's perspective on the above-referred areas with the insights from *E1.* However, we say "*to some extent*" as *E1* could only provide significant outsider's insight on the area of the data marketplaces as he is well-versed with the concepts dealt in that area. With respect to the area of MPC technology, *E1* did not have proficient expertise but he contributed to the extent of his knowledge which although was not significant, but still a considerable contribution which helped towards the refinement of few concepts in the area of MPC Technology. This problem was not experienced in the area of threat modelling as most of the insights obtained here were from the perspective of an outsider namely, *E1* and *E3.*

6.3 Procedure

The expert interviews were semi-structured interviews and were conducted over Skype. Prior to the interview, the experts were directed to be familiar with the concepts described in the attached document and were asked to have a copy of the same document with them so that it is easier for them to follow when the concepts are referred during the interview. Before starting the interview, the consent of the expert was taken verbally to record, transcribe and use the insights from the interview in our research. After taking the consent, the interviews were recorded over an android phone. Once the recording started, the same consent was taken verbally again so that the consent

was also on record. After this, it was asked to confirm if the expert had a chance to familiarise himself/herself with the concepts of the relevant artefacts prior to the interview. Unfortunately, all the participating experts did not study the document; instead, they just skim read the document. This would have been a setback as we were validating concrete concepts which needed prior understanding rather than just asking for open opinions. However, we had devised a solution for this problem. An overview of the research and the relevant artefacts was verbally described for almost 10 minutes before starting the actual interview. This solution was further solidified by verbally explaining each concept being dealt before asking the corresponding question. The interview was carried out by asking the previously-prepared semi-structured questions, the follow-up questions and the follow-up discussion which went on until a comprehensive understanding was reached on each concept. The interview questions can be found the interview transcripts in Appendix A**.** At the end of the interview, the experts were thanked for their participation and the skype call was ended. Table 21 shows the topics on which the insights were provided by each expert.

***Table 21**: Topics validated by each Expert*

Topic/Expert	*E1*	*E2*	*E3*	*E4*
***T1:** Data Marketplace Platform Designs*	✓	✓	-	✓
***T2**: Functional Requirements of the Data Marketplace Platform*	✓	✓	-	-
***T3**: Customers of the Data Marketplace Platform*	✓	✓	-	-
***T4**: Functional Components of the Data Marketplace Platform*	✓	✓	-	-
***T5:** HLA Framework*	✓	-	-	-
***T6:** Perception of MPC Technology*	✓	✓		✓
***T7:** MPC Incorporation into the Data Marketplace Platform*	✓	✓		✓
***T8:** HLTM framework.*	✓	-	✓	-
***T9:** Threat Landscape of the Data Marketplaces*	✓	-	✓	✓
***T10:** Effect of MPC Incorporation on the Threat Landscape of the Data Marketplace Platform*	✓	-	-	✓

The criteria for stopping the data collection and analysis was initially considered to be *theoretical saturation* where no new information emerges from the successive interviews (Sekaran & Bougie, 2013). However, because of the time constraint of the research, we then subjected ourselves to a deadline until which interviews will be conducted. The deadline was decided to be ***1 August 2019*** which provided us with exactly one month for the qualitative data analysis and report writing before submitting for the Green Light Meeting which was scheduled on ***6 September 2019.*** Fortunately, every concept associate with every artefact was at least validated once in the interviews. However, theoretical saturation was not reached, and only the results obtained so far are presented in this thesis.

6.4 Analysis

Each interview was transcribed, and insights were understood right after the interviews; so that the insights could be incorporated to refine the concepts and further, these refined concepts can be referred in the further interviews. Accordingly, the insights were understood and appropriately addressed in the forthcoming interviews. This helped in deepening the understanding of the concepts as the number of the interviews progressed. However, the formal qualitative data analysis was carried out after all the 4 interviews were done. The qualitative data analysis was carried out with the 3 traditional steps: *Data Reduction, Data Display* and *Drawing Conclusions*. The procedure followed for the analysis is further explained in this section.

In *Data Reduction*, since we already had the initial list of categories and codes of each topic, we moved directly to the second phase of coding, ***Analysis phase: Axial Coding*** (de Reuver, 2019b). Here, we mapped the statements and insights from the interview transcripts to their appropriate categories and codes. Subsequently, with this mapping, we analysed and carried out the refinement, updation and modification of the concepts of all the categories and codes. After this process, with the data that is left unrelated to the existing codes, new codes were created for these unmapped insights and were assigned to their appropriate categories and topics. The whole data reduction was done manually using a data log book where we documented the constant comparison between the interview transcripts and the then list of categories and codes. No software was used to carry out the data reduction. As a result, there was no illustrative way to visualise the data reduction and hence, data reduction was decided to be represented in a qualitative way (basically, in ***words***) as opposed to the traditional ways of data visualisation (like matrix, timeline, networks, actor network, process (de Reuver, 2019b)). However, we illustrate the categories and codes in either of the lists (initial and updated) are illustrated before and after the analysis in each topic in the form of *figures, tables, lists, hierarchies* or just *textual descriptions*

Moving on, the ***Results & Analysis*** section was written for each topic signifying the *Data Reduction & Data Display* step of qualitative data analysis. Here, the data mapped to the appropriate concepts i.e. the data reduction and data display are *represented in a qualitative way* by relating it to the respective expert; If the resulting code relates to the concepts already associated with the initial set of categories and codes, they are represented in *Italic* font while the newly emerged concepts and their codes are displayed in ***bold-face*** font; both contributed towards generating the updated list of categories and codes. Following this, we wrote the ***Drawing Conclusions*** section for each topic signifying the last step of the same name of the qualitative data analysis. These sections collectively contain the updated iterations of all the concepts refined, updated or modified after incorporating either the quoted insights, further analyses or further implications to obtain a more deeply valid concepts of each topic. These represent the updated list of categories and codes associated with the concepts of each topic. Finally, using these updated iterations of the concepts, all the sub-research questions associated with the research task *RT5* are answered which provide relatively more valid contribution in answering the main research question.

There was an anomaly with one of the topics, *T9* for which all the initial list of categories and codes were totally disregarded and discarded during the analysis process. This issue will be dealt with proper reasoning in Chapter 7 when addressing the topic, *T9: High-Level Threats associated with the Data Marketplaces.* Later, we generated a new list of categories and codes from the interview transcripts alone by carrying out the first phase of data reduction which is, ***Exploration phase: Open Coding*** (de Reuver, 2019b)***.*** Here, the information related to *T9* in the transcripts was traversed repeatedly to obtain observations; which was later reduced to obtain the list of categories and codes.

Because of the shortage of the number of interviews and the time constraint, we could not continue the analysis further. As a result, we did not get a chance to carry out the last phase of data reduction, ***Reduction phase: Selective Coding*** through which we could have validated the relationships between the categories and codes with a greater number of interviews. This could have potentially helped us to reach *theoretical saturation* and obtain an ultimate list of categories and codes in each

topic with most-refined theoretical concepts emerging from a single core category (our research objective). Through this, we could have ended up with the relatively most-valid answer to our main research question. Unfortunately, because of the reasons mentioned here, it was not possible to pursue this, resulting as a major limitation to our research. This is also the reason we refer to the list of categories and codes obtained after the analysis as updated list but not final list.

This brings us to the end of the completion of the research task, *RT4* which is to formulate the methodology to carry out the validation. Using the methodology described here, we performed the research and the corresponding results, analyses and further conclusions are discussed in the Chapter 7 in an extensive detail.

7

Results and Analyses

This chapter reports the findings of the ***Validation phase*** carried out as part of the final research task, *RT5: To validate the artefacts from the conceptualisation phase*, which entailed to validate the conceptual models, *Artefacts 1.0* and their subsequent theoretical concepts through *qualitative data analysis.* Following this, the refined conceptual models, ***Artefacts 2.0*** are obtained along with their more valid and refined theoretical concepts using which the sub-research questions, *SQ8, SQ9, SQ10, SQ11, SQ12, SQ13, SQ14* and *SQ15* can be answered in Chapter 8.

The sections of this chapter are divided according to the 4 research foci. The chapter comprises of the results, analyses and further conclusions on the theoretical concepts of the 10 topics. The findings are reported in the template of the steps associated with qualitative data analysis: *Data Reduction, Data Display and Drawing Conclusions*. These 3 steps are represented in 2 subsections: ***Results & Analyses*** (reflecting data reduction and data display) and ***Drawing Conclusions***; and these subsections are written either for each topic (*T*) or the whole research focus (*RF*); whichever is applicable. But every theoretical concept associated with every topic is validated in this chapter. The basic template used here for representing the process of qualitative data analysis is in the format as illustrated in Figure 17 and the same is followed for each topic.

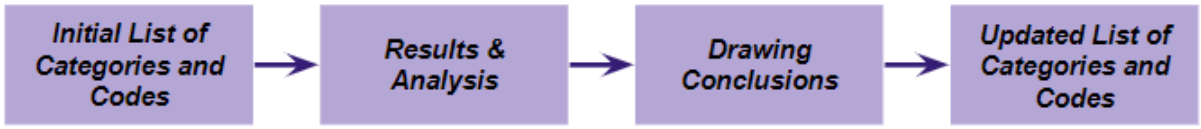


Figure 17: *Template for representing the Qualitative Data Analysis in each Topic*

7.1 *RF1*: Validation of Pre-MPC Data Marketplace Platform 1.0

The topics and corresponding theoretical concepts associated with the research focus, *RF1* are validated here. The artefact under consideration here is the *Pre-MPC Data Marketplace Platform 1.0* built in Chapter 2. The following 6 topics under *RF1* are validated in the upcoming subsections.

- ***T1:*** *Data Marketplace Platform Designs*
- ***T2:*** *Functional Requirements of the Data Marketplace Platform*
- ***T3***: *Customers of the Data Marketplace Platform*
- ***T4:*** *Functional Components of the Data Marketplace Platform*
- ***T5:*** *HLA Framework*

7.1.1 *T1*: Data Marketplace Platform Designs

In Chapter 2, the potential platform designs of the data marketplaces were discussed as proposed by Koutroumpis, Leiponen, & Thomas (2017) which involved, ***centralised, decentralised*** and ***collective platforms***. However, these were predictive conceptualisations proposed based on the economic perspective of the institutional requirements: *boundary conditions*, *rules* and *monitoring mechanism*. In addition to the functional requirements from subsection 2.5.1, these conceptualisations do not consider other design aspects of data marketplaces like that architectural aspects, business processes, enabling technologies like homomorphic encryption, multi-party computation et cetera and their maturity to implement into the data marketplaces. The designs were just theoretical frameworks and hence, they do not reflect the real-life platform designs of the data marketplaces. For this reason, this topic, *T1: Data Marketplace Platform Designs* was considered for an exploratory study under the hope to enhance their understanding with expert insights.

The theoretical concepts associated with this topic, *were* analysed by relating them to the insights of experts *E1* and *E2*. The initial list of codes in this topic derived from Chapter 2 were:

- ***T1:*** *Data Marketplace Platform Designs*
  - ***C1:*** *Centralized Platform*
  - ***C2:*** *Decentralised Platform*
  - ***C3:*** *Collective Platform*

7.1.1.1 Results & Analysis

When asked about the real-life data marketplaces and their platform designs, *E1* responded by saying "*the term, data marketplaces, is a bit overused*" and suggested that even single domain data provider who provisions data over a cloud also calls himself a data marketplace. This insight is in similar lines with our criticism towards the systematic survey of Schomm et al. (2013) which includes even data vendors in their survey of data marketplaces and subsequently, suggests the focal data marketplaces (multilateral B2B data marketplace) in this research as just one of the categories in their classification. *E1* suggests that the ideal design of a data marketplace is to have a "*distributed system similar to Internet Exchange*" where anybody can hook up to the marketplace and carry out data exchange with anybody. We name this code as ***truly many-to-many data marketplaces.*** *E1* claims that this kind of design is theoretically possible and is being worked on. However, the execution of such a marketplace is complex and the idea is not realised yet owing to many reasons. Speaking on the real-life data marketplaces, *E1* suggested that the actual data marketplaces that do exist are formed in the lines of a ***consortium*** where "*parties within an industry come together to figure out a way to share data such that it is profitable for all the parties*" involved. Following this, the parties figure out a ***use-case*** to generate value out of data and create an architecture of a *data marketplace for that specific use-case* with fixed actors and fixed processes***.*** Furthermore, *E1* touches upon the possibility of *centralised* and *decentralised* data marketplaces in the same meaning as our initial codes; which is based on where the physical data resides. He says that *decentralised design* is operational with the help of a "*key management system*". In this case, a data provider holds the data and provisions his data with the help of public key encryption where the dedicated data consumer holds the private key and gets access to that data. Since this involves a requirement for governance to manage the public and private keys, this kind of model would not realise *truly many-to-many data marketplaces* where governance is complex because of its true many-to-many nature. However, in a closed consortium with fixed limited members, the governance of key management and subsequent data transactions is feasible. *E1* suggests another way of materialising *decentralised* design is by putting the data on ***blockchain*** "*but it is not feasible yet for real-life application*".

When asked *E2* about the platform designs of data marketplaces, he reflects on a ***truly many-to-many data marketplace*** that it is not possible to realise it for various reasons. The *absence of data sharing culture* is one of them. *E2* suggests that in a practical sense, the realisation of data marketplaces is driven by the ***use-case*** through which the data is utilised. Once the *use-case* is developed, data can be brought onto the platform easily from the data owner. However, *E2* also suggests that it is difficult to foresee a *use-case* without the availability of the data and its details. *E2* relates to this as a chicken-egg problem. However, *E2* also discusses the possibility of a platform where an innovator who innovates the use case can search for the appropriate data on that platform. On this kind of data marketplace platform, the innovator can also browse through the data catalogue using the metadata provided on the platform by the data owners and if interesting data is found, can innovate a *use-case*. *E2* reflects that the former case is more likely than the latter one. *E2* calls the latter kind of data marketplaces as ***"serendipity model"***. *E4* also echoes the *serendipity model* by referring it as a platform where the companies who have data and the companies who want to run statistics on such data can find each other.

7.1.1.2 Drawing Conclusions

Combining the above-discussed insights into the initial codes of *T1*, we built a taxonomy for the platform designs of the many-to-many data marketplace platforms reflecting the expert insights; thereby, replacing the previous classification. The taxonomy represents the updated list of categories and codes of *T1*. Broadly, the taxonomy consists of 2 categories of platform designs based on where the data resides: *Centralised* and *Decentralised**.***

- In ***centralised*** design, the data is transferred from the data owner and stored on the platform and the data consumer finds the data on the platform and downloads it for his/her use. Since the owner loses the control over the data, only low value data like open data is transacted through such platforms.
- In ***decentralised*** platforms, the data resides at the data owner's site and is accessed only by dedicated data consumer or data aggregator through some encrypted channel. Since the data owner has the control over his data and the data consumer is allowed to access that data over contractual obligation facilitated by the platform, high value data can be transacted on such platforms. Further, in *decentralised* design, there can be 2 variants based on the design specification of the data marketplaces related to its ecosystem design, technological architecture design et cetera. The variants are *truly many-to-many data marketplace, blockchain based data marketplace* and *closed consortium data marketplace**.***
    - The ***truly many-to-many data marketplace*** is the ideal design where anybody can log into the platform and provision their data to anybody else on the platform, as suggested by *E1* and reflected by *E2*. This is the end goal for the species of data marketplaces which is feasible only in time when other factors like technological maturation, data sharing culture et cetera come together.
    - The species of the ***blockchain based data marketplace*** is straightforward as suggested by *E1* where the data transaction happens through a block chain. The data owner uploads his data to the blockchain and the data consumer access the data on the blockchain. Meanwhile, the blockchain monitors all the activities being carried out on that data which is stored, and any anomaly will be reported. This design relates to the *decentralised platform* as suggested by Koutroumpis et al. (2017). The design is being worked upon and is expected to materialise once the blockchain technology attains mainstream maturity which is not very far in the future.
    - The ***closed consortium data marketplace*** are the data marketplaces formed by parties within an industry to share data among each other. This variant is similar to the *collective platforms* as suggested by Koutroumpis et al. (2017) which already operate in the real world. Furthermore, in *closed consortium data marketplaces**,*** we have included 2 more

subcategories based on the business process associated with them. They are: *use-case based data marketplace* and *serendipity model data marketplace*.

- In a ***use-case based data marketplace***, a fixed number of data actors come together to form an architecture driven by a specific use-case which defines the business process of the data marketplace. In this variant, the business process and the roles of the actors in the architecture will be fixed while the companies representing the actors can plug-in as and when necessary to transact the data, satisfying the many-to-many criteria. The data marketplace proposed by Roman & Stefano (2016) can be attributed as an example for this variant. This design was seconded by *E1* and *E2* as the most-likely and a realistic design for a data marketplace as this design practically exist in operation in the real world.
- The ***serendipity model data marketplace*** is a platform where the data owners within the consortium can showcase their data in the form of metadata for the potential data consumers in need of that data and consequently, form a relationship and share data among only each other in an ad-hoc sort of way with a communication channel. Here, other data actors like data managers and data aggregators also showcase their services on the platform to find data partners. This design is more flexible with no business process fixed for the data trading but is formed when the data actors find each other with their data and corresponding use-case for the utilisation of that data.

The taxonomy reflects the final list of the categories and codes of *T1* and is illustrated in Figure 18 in the form of a hierarchy. This serves as an update to the classification of Koutroumpis et al. (2017) and also extends the category of ***Data Market Place*** in the classification of Schomm et al. (2013).

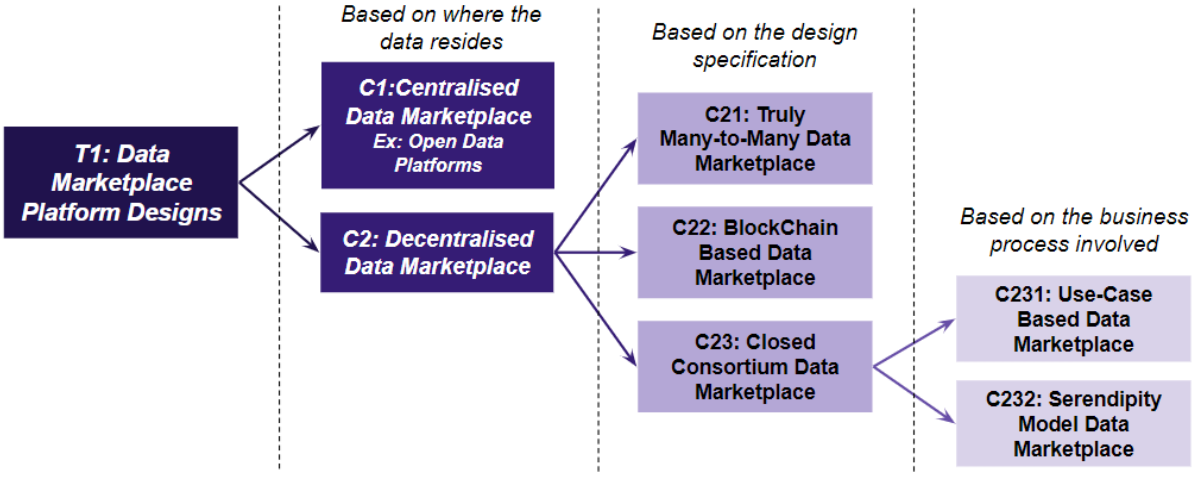


***Figure 18**: Data Marketplace Platform Designs Taxonomy 2.0*

After all these different designs of data marketplace platforms were established, it was deduced that our focal data marketplace (multilateral B2B data marketplace) as illustrated using *Pre-MPC Data Marketplace Platform 1.0* from Chapter 2, related to the *Serendipity Model variant* in the *closed consortium* category from the taxonomy. We combined this insight and refined our scope. The resulting species of the data marketplaces which was focussed from then on was ***Many-to-Many B2B Decentralised Serendipity Model data marketplaces.*** In the rest of the thesis, when we refer the term data marketplace, we mean this species. The reason for doing so was that through our knowledge from the study on data marketplace so far, it was deduced that this species represented the most generic form of a data marketplace which coincided with the one referred in *RT1.*

7.1.2 *T2*: Functional Requirements of the Data Marketplace Platform

The functional requirements were dealt comprehensively in Chapter 2. However, the actual meaning of these requirements was needed to be understood to check if it reflects the same as our interpretation. Hence, the topic, *T2: Functional Requirements of the Data Marketplace Platform* was included as an exploratory study to be validated and refined.

The theoretical concepts associated with this topic were analysed by relating them to the insights of experts *E1* and *E2*. The initial list of categories and codes in this topic derived from Chapter 2 were:

- ***T2:*** *Functional Requirements of the Data Marketplace Platform*
  - ***C1:*** *Boundary conditions*
  - ***C2:*** *Data Provenance*
    - ***C21:*** *Data Lineage*
    - ***C22:*** *Change of Ownership of Data Point*
  - ***C3:*** *Data Governance*
    - ***C31:*** *Management of Data*
    - ***C32:*** *Data Exchange Traceability*
    - ***C33:*** *Data Usage*
  - ***C4:*** *Data Economy*
    - ***C41:*** Revenue
  - ***C5:*** *Data Sovereignty*
    - ***C51:*** Handling Permissions
    - ***C52:*** Usage Restriction
    - ***C53:*** Data Contacts
  - ***C6:*** *Secure Data Exchange*
  - ***C7:*** *Data Exchange Platform*.

7.1.2.1 Results & Analysis

When discussing about the general requirements for a data marketplace platform, *E1* suggested that the starting point here is having a ***governance*** model. He expands on enforcing *governance* as an "*an authority who manages all the parties and activities*" on the data marketplace platform. One of the activities involves handling the ***legal aspects*** comprising of contracts which "*contain the terms of what can be shared with who, which data can be shared using which algorithm, what computing functions can be done in this algorithm, timeframes, quality if the data et cetera*". Furthermore, *E1* adds another requirement associated with the *governance* which is ***trust mechanism***. *E1* says *trust mechanism* is enforced again with independent authorities like ***Certification Authority***, ***Auditing Authority*** et cetera. These authorities with their activities bring about the trust on the data marketplaces in an indirect and intangible way. *E1* reflected on our assumption of enforcing data governance just through technology alone that technology can "*kind of enforce the governance but there is no way to restrict technologically when someone among the parties can just copy the data and run away with it*". *E1* says that something like this can only be tackled from the legal angle, with an authority and not from the technological angle. Basically, *E1* says "*the complex thing is to find a right coordination between the technology and legal aspects to have a complementary effect*". Basically, the requirements can only be enforced if both the aspects of technology and legal angle are in place and it cannot be done by just one of them. *E2* did not touch up on these issues and went right about reflecting on the functional requirements we had compiled from the literature.

Moving on to reflecting on the initial set of functional requirements from Chapter 2, *E1* and *E2* had several comments.

- ***Boundary conditions:*** This requirement is referred as necessary by *E1* and stated that it also depends on how it is implemented, and it is a topic in itself to explore. *E1* also suggested that this requirement is part of the ***governance*** aspect. *E2* reflects that our phrasing of *boundary conditions* is good and states that it "*is required*".

- ***Data Provenance:*** *E1* stated data provenance as "*an important aspect*" and suggested that this requirement is enforced through auditing of the transactions. The audit trail gives the *data provenance.* Both *E1* and *E2* had problem with our phrasing in the description of the concept of data provenance. *E1* suggested that the phrase "*change of ownership of data points*" used in the description of data provenance is not clear and it should be defined precisely given that it can have different implications. *E1* says technologically, the *ownership of data* can be defined in terms of *ownership of private key* to access the data in which case, the switching of private keys signifies the *change of ownership of data*. A ***key management*** component will come in place there as part of the identity management. However, *E1* says there is risk involved here if it is done without any governance as in that case, even if the change of ownership of data, the owner can have a copy of the same data and he can sell it to other party. So, the governance model should take care of this aspect such that the ownership change happens according to the terms in the ***contract.*** *E2* also reflects on our phrasing of *change of ownership of data* and discards the concept saying that in data markets, ownership of data does not exist and what exists are ***licences***. *E2* says that "*there is no process involved where the change of ownership happens".* Furthermore, *E2* clarifies the meaning of ***data lineage*** by saying *"it is the transformation of the data from its origin to the current state*" and ***data usage*** by relating it to "who has access to the data, who accessed it and whether they accessed it or not. These 2 concepts are the constituents of the requirement of *data provenance*.

- ***Data Governance:*** *E1* states *data governance* as "*the most important requirement which establishes the legitimacy of the data marketplace*". It is enforced through an authority actor who oversees all the operations on the marketplace. However, it can also be enforced through technology, but it depends on the architecture of the marketplace. *E1* gives an example where an ***authority*** facilitates the ***contract*** of data exchange among data actors. *E1* says that contracts define the business process of using technology to carry out *data exchange*. So, *E1* says the requirement of *secure data exchange* is also governed by the governing authority, stressing that an authority actor is necessary for ***governance,*** and that *secure data exchange* is a part of *governance* requirement. On the other hand, *E2* reflects on our description of data governance and states that "it is a combination of the *secure data exchange, data sovereignty* and *data provenance*". This statement relates to what *E1* stated earlier that *governance* involves managing all the activities of the data marketplace.

- ***Data Economy:*** *E2* agreed with our description of *data economy* saying that it is fine to be a requirement. *E1* did not touch on this.

- ***Data Sovereignty:*** *E1* thinks that it is true that data sovereignty can be enforced but it depends on the design. He says, in a *centralised* design, the central authority has the control over data and sovereignty here means that the data owner trusts the central authority to do what the owners asks him to do. But it can be truly enforced in *decentralised* design by keeping the data on blockchain where the data owner can control it. However, if the data is copied, then *data sovereignty* is lost. But since there is no real life blockchain application on this yet, *E1* says this is a direction to investigate. *E2* thinks of *data sovereignty* as a requirement to be fine. However, *E2* reflects again on our phrasing in the description of *data sovereignty* that it is not about protecting the *legality of the data* as "*the data is either legal or illegal*". *E2* suggests that *data sovereignty* is basically having ***control*** over who uses the data.

- ***Secure Data Exchange:*** *E1 and E2 were fine with our description of the secure data exchange and it being a requirement.* However, *E1* had a concern relating to this subject that "*once the consumer gets the data, nothing stops him from doing whatever he wants with the data*". *E1* says this issue as the more pressing issue than an external entity intercepting the transacted data.

E2 had a phrasing issue over the consistency of the term ***data actors*** as we used inconsistently with the terms "*data customer*", "*data subjects*".

- ***Data Exchange Platform:*** *E2* found this requirement to be redundant as it is the complementary requirement of rest of the requirements.

Reflecting on overall of requirements, *E1* remarked that ***governance*** is the fundamental requirement and the rest of the requirements is dictated by the use-case and the architecture of the data marketplace platform. On the other hand, *E2* reflected that the requirements are "*reasonable*" to have for a data marketplace platform; while also suggested that these requirements are "*exhaustive in the sense that they are generic*" and the requirements cover all the bases relevant to a data marketplace.

7.1.2.2 Drawing Conclusions

Although the theoretical concepts associated with our requirements were only from technological standpoint, we realised we should include the non-technological aspects to have a comprehensive understanding of the requirements. This was also recommended by expert, *E1* as he said it is not possible just with technology alone, but we need a non-technological governing authority to effectively achieve the fundamental functioning of the data marketplaces. Furthermore, the interpretation of each requirement was also validated and are refined here as applicable to reflect the credible expert insights. Furthermore, after the analysis, it was deduced that the functional requirements should provide objective description of the requirements applicable to the data marketplace platforms. As a result, when describing the functional requirements here, we omitted from the description, the examples of how they are enforced by the data marketplaces as they are implementation-dependent but not objective information.

- ***Boundary Conditions:*** The description of the boundary conditions remain the same as before which is, "*Strict boundary conditions help in authorising only the legitimate costumers willing to share or buy data. This helps in safeguarding the data from unauthorised access*".

- ***Data Provenance***: This requirement undergoes changes in its description where we omit the phrase "*change of ownership*" as the concept was disregarded by *E2.* Although, considered as a possibility by *E1,* it is never observed to be in practice. The concept that does exist in data marketplace is the concept of *licenses*. Practically, the data owner always owns the data and, he provisions the data to the data consumer who can use it according to the terms agreed in the licensing contract. So, we change the phrase "*change of ownership*" to "*data usage*" which is actually in lines with the meaning of data provenance. So, the description changes to "*Data Provenance is a requirement to track and document the* ***data lineage*** *and* ***data usage****. Data lineage refers to the transformation of the data from its original state to the current state (different versions). Data usage is focussed on who has the access to the data, who accessed it and if they accessed or not".* The metadata aspect is omitted from the description here as the enforcement of data provenance is implementation dependant and is more a part of functional components which deals with features like that of metadata.

- ***Data Economy:*** This requirement remains the same too which "*reflects the business purpose of the data marketplace platform which is to generate revenue stream for itself through its services*". However, we have excluded the information about its way of implementing.

- ***Data Sovereignty:*** After discarding the phrasing of "*legality of data*", this requirement can be described as *a mechanism expected for the data marketplace platform to support for the data owner to have control over his data and its usage".* The examples of how it is implemented like that of *handling permissions, laying restrictions on usage or by provisioning the data via blockchain* et cetera are omitted.

- ***Secure Data Exchange:*** There is no change in interpretation of this requirement. Its description remains the same as "*the most fundamental aspect of the data marketplace platform which is carry out the data exchange between the data actors in the most secure way".*

The requirement of *Data Exchange Platform* is removed from the list as it is declared as redundant. Moving on, the new requirements that evolved from the expert insights and further analysis were also included into the list. These are described as follows,

- ***Marketplace Platform***: This requirement is a transformed version of the *Data Exchange Platform* which basically deals with the platform aspects of the data marketplaces which is obviously a fundamental requirement for a data marketplace platform. This requirement is described as "*the requirement of platform features like match-making between the participants; and the marketplace features like cataloguing, curation, e-commerce mechanism, recommendations et cetera*". This description makes way more sense than the previous one of *Data Exchange Platform* and hence, also makes it different.

- ***Legal Management:*** This requirement is for the data marketplace platform to handle the legal aspects of data trading like *contract management, license management, litigation* etc. This requirement is enforced by a human actor and not by technology.

- ***Trust Mechanism:*** This is also a non-technological requirement enforced by a different kind of human actor which is more like an independent authority, for example, *Certification Authority, Auditing Authority* et cetera; who through their operations, create trust for the data actors to participate in data trading over the data marketplace platform.

The requirements of Legal Management and Trust Mechanism can not necessarily be enforced by the data marketplace platform itself but can be done on an ad-hoc basis by external entities which possess expertise of specific issues like *Certification, Auditing, Legal Counsel* et cetera. Furthermore, *Legal Management* and *Trust Mechanism* can currently be enforced purely by authority actors on the data marketplace platform; while the rest depend on their implementation consisting of a coordinated effort both technology and actors. However, cutting-edge technologies like BlockChain, Multi-Party Computation (MPC), Homomorphic Encryption et cetera can enable the data trading technologically alone without any human actor. But this is just a claim as the said-technologies have not achieved the desired level of sophistication to be applied in real-life cases. Evidently, investigating this claim is part of our research problem but we are only doing it with respect to MPC technology.

Moving on, the above list represents the updated functional requirements and we categorise all of these under a core category reflecting the most fundamental requirement for a data marketplace platform to satisfy which is, ***Governance***. *Governance* can be described as the requirement of a mechanism which oversees all the activities on the data marketplace. As specified by *E1*, it can only be enforced by the right coordination between the human actor and the technology; however, difficult with one of them alone. Furthermore, subcategories were created for this core category. Since the requirements of *Data Provenance, Data Economy, Data Sovereignty* and *Secure Data Exchange* relate to the overseeing of the activities associated with data, we group these requirements under the subcategory, ***Data Governance.*** On the other hand, we group *Boundary Conditions, Marketplace Platform, Legal Management* and *Trust Mechanism* under the category of ***Marketplace Governance*** as they comprise of overseeing the activities specifically of the marketplace aspect. The updated list of categories and codes reflecting the refined functional requirements is listed below and is illustrated in Figure 19 in the form a hierarchy under the core category of *Governance*.

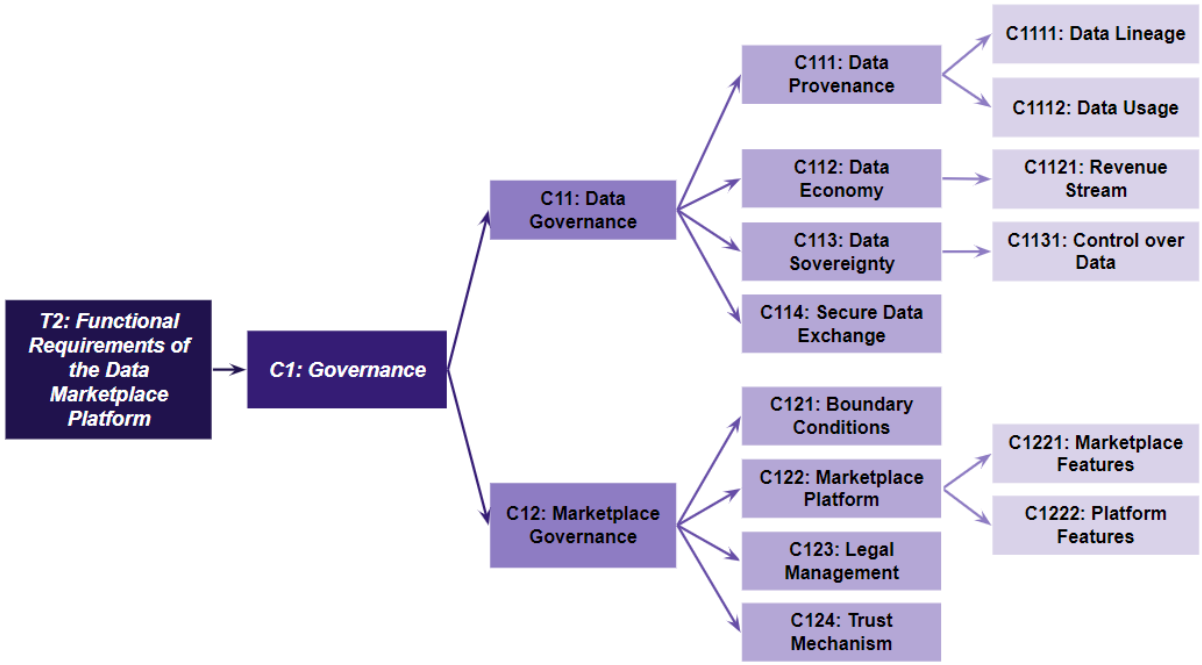


Figure 19: *Functional Requirements 2.0 of the Data marketplace Platform*

7.1.3 *T3*: Customers of the Data Marketplace Platform

The list of customers dealt in Chapter 2 was not an exhaustive list. Hence, the topic, *T3: Customers of the Data Marketplace Platform* was included as an exploratory study so that the list can validated and updated to have a more exhaustive list. The theoretical concepts associated with this topic were analysed by relating them to the insights of experts *E1* and *E2*. The initial list of categories and codes in this topic derived from Chapter 2 were:

- ***T3:*** *Customers of the Data Marketplace Platform*
  - ***C1:*** *Data Providers*
    - ***C11:*** *Data Collectors*
    - ***C12:*** *Data Managers*
    - ***C13:*** *Data Aggregators*
  - ***C2:*** *Data Consumers*

7.1.3.1 Results & Analysis

In terms of the actors, *E1* stresses the significance of an ***authority*** who is according to him, very crucial for the governance of the data marketplace. *E1* also mentions different *authorities* which carry out different functions in the data marketplace like *Certification Authority, Auditing Authority* et cetera. *E2* suggests on maintaining the consistency of the terminology in the descriptions with what is used in the industry like ***data actors*** instead of *data subjects*.

7.1.3.2 Drawing Conclusions

With the refinement of our focal data marketplace to *many-to-many B2B serendipity model*, the updated list of functional requirements combined with the expert insights, we decided to include further actors into the architecture who are not customers but play a crucial role for the functioning

of the data marketplaces. As a result, we renamed the core category from "*Customers*" to ***Actors of the Data Marketplace Ecosystem*** to reflect the ecosystem view of the data marketplaces.

To maintain the terminology consistent with the industry usage, we modified the core categories of Data Providers and Data Consumers into a single category named as ***Data Actors*** which reflect the customer definition from the initial list. Further sub-categories were added; namely, ***Data Supply*** side and ***Data Demand*** side which are consistent with the industry usage. In the *Supply side*, we included the actors who supply data and data related services on the data marketplace platform; basically, *Data Owners, Data Managers, Data Aggregators* and even third-party data analysis service providers. On the *Demand side*, we put *Data Consumers.* All the actors retain their previous interpretations from the initial list in the sense that they use the platform services for their benefit.

Apart from these *data actors*, we also included the actors who enable the data marketplaces like the *authority services* as stressed by the experts. We termed these actors as ***Marketplace Enabling Actors***. These actors represent the ***network*** aspects where the criteria for the actors expands from the usage and non-usage of the services to the creation and capture of ***value*** in the data marketplace system. We further divided the enable actors into 2 categories:

- ***Marketplace Provider:*** This actor is the central authority who provides the data marketplace service by hosting and managing all the services and operations on a technological platform. This is an organization whose business model is to provide the data marketplace service and enforces the requirement of *Governance* by implementing the business processes using either technology or just human actors.
- ***Independent Service Providers:*** These actors are independent actors who provide services to enable the data marketplaces as and when necessary. The services can range from technological services like infrastructure provider to non-technological services like certification, auditing, legal counsel et cetera. Mostly, these actors enforce the non-technological requirements like *Legal Management, Trust Mechanism* et cetera.

The updated list of the categories and codes reflecting the actors in the data marketplace ecosystem is illustrated in Figure 20.

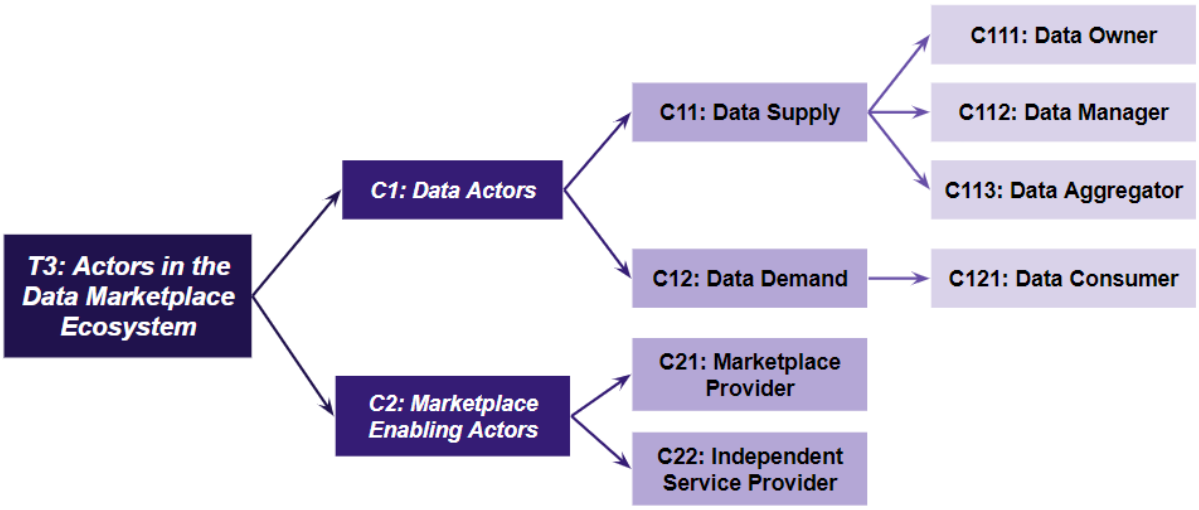


Figure 20*: Actors 2.0 in the Data Marketplace Ecosystem*

7.1.4 *T4*: Functional Components of the Data Marketplace Platform

The functional components developed in chapter 2 were the result of our desk research and hence, they were needed to be validated with expert insights to refine into more valid components which reflect the features which were empirically expected to be on the data marketplace platform. Hence, the inclusion of the topic, *T4: Functional Components of the Data Marketplace Platform* for an exploratory study where the related theoretical concepts were analysed by relating them to the insights of the experts, *E1* and *E2*. The initial list of categories and codes in this topic derived from Chapter 2 were:

- ***T4:*** *Functional Components of the Data Marketplace Platform*
  - ***C1:*** *Identity Management*
    - ***C11:*** *Features*
      - ***C111:*** *Induction*
      - ***C112:*** *Authentication*
      - ***C113:*** *Authorization*
    - ***C12:*** *Enforced Requirements*
      - ***C121:*** *Boundary Conditions*
  - ***C2:*** *Broker Service*
    - ***C21:*** *Features*
      - ***C211:*** *Data Management Services*
        - ***C2111:*** *Data Cataloguing*
        - ***C2112:*** *Data Marketplace Curation*
          - ***C21121:*** *Data Categorisation*
          - ***C21122:*** *Data Tagging*
        - ***C2113:*** *Data Tracking*
          - ***C21131:*** *Data Lineage Tracking*
          - ***C21132:*** *Data Usage Tracking*
      - ***C212:*** *User Interaction Services*
    - ***C22:*** *Enforced Requirements*
      - ***C221:*** *Data Exchange Platform*
      - ***C222:*** *Data Governance*
      - ***C223:*** *Data Provenance*
      - ***C224:*** *Data Economy*
  - ***C3:*** *Clearing House*
    - ***C31:*** *Features*
      - ***C311:*** *Transaction Repository*
    - ***C32:*** *Enforced Requirements*
      - ***C321:*** *Data Provenance*
  - ***C4:*** *Data Inventory*
    - ***C41:*** *Features*
      - ***C411:*** *(Meta)Data Storage*
    - ***C42:*** *Enforced Requirements*
      - ***C421:*** *Data Governance*
      - ***C422:*** *Data Sovereignty*
  - ***C5:*** *Data Exchange Service*
    - ***C51:*** *Features*
      - ***C511:*** *Data Exchange Mechanism*
    - ***C52:*** *Enforced Requirements*
      - ***C521:*** *Secure Data Exchange*
  - ***C6:*** *Data Analytics Service.*
    - ***C61:*** *Features*
      - ***C611:*** *Data Analytics Tools*

- ***C62:*** *Enforced Requirements*
  - ***C621:*** *Data Economy*

7.1.4.1 Results & Analysis

The experts reflected on our conceptualised components one by one and their comments on each component is discussed below.

- ***Identity Management*****:** *E1* suggests that if we are dealing with a decentralised data marketplace, then data exchange happens through an encrypted channel involving public key encryption. Since the data marketplace is responsible for the data exchange, it should generally have a ***key management system*** to manage keys and in turn, the communication channels. Apart from this, both the experts *E1* and *E2* were "*fine*" with our features of *Induction, Authentication* and *Authorisation.*

- ***Broker Service:*** *E1* validates that we "*need catalogues of data objects and metadata of each object to describe the data that is being showcased on the data marketplace platform*". Furthermore, *E1* says the management of *physical data* is also done by the *broker*. *E2* says that in Data Market Austria, they separate *metadata* and the *data*; the *metadata* is centralised and is completely relied on by the *broker service*. *E2* further reflects on the feature of *data tracking* and reflects that only *data lineage* can be part of the *data tracking* feature, while the *data usage* is more applicable in the context of ***transaction management***. Other than that, both experts were okay with the rest of features of the *broker service.*

- ***Clearing House:*** *E1* stressed that since the *transactions* are needed for auditing purpose, the management of *transactions* is crucial for the data marketplace. Hence, *E1* suggested that this should be implemented using some tamperproof database or blockchain ledger. *E2* reflected that the feature of *data usage tracking* should be integral to the *clearing house* component.

- ***Data Inventory:*** *E1* shared his scepticism on how the decentralised design can be materialised. *E1* explained a possibility with the help of *key management system*. *E1* said that *data can be on the provider's site* and if we want to compute something on the data, we can have a container with an algorithm which needs to decrypt data. So, *E1* said it goes to *key management* again to manage the credentials of the data. *E1* also stated that however, the container with the algorithm can copy the data for itself which breaks the security. So, *E1* said we need ***governance model*** to manage this situation. On the other hand, *E2* pointed out the *data management* feature is part of *broker service* and it does not make sense to have a *data inventory* component. Hence, *E2* suggested that we can exclude *data inventory* component. *E2* also remarked that it does not matter where the physical data resides as it can be stored on a distributed system and its access can be managed by *broker service.*

- ***Data Exchange Service*****:** *E1* did not have any comments here except for perceiving it just as a ***communication channel*** enabling *secure data exchange*. However, *E1* reflected on different aspects of designing the business process of the data sharing among the actors; some of which were: how the infrastructure of the data sharing is designed, whether the parties have preferences there, how the data access is provided, whether through algorithm or a container. *E1* suggested all these aspects to be related to *secure data exchange* and hence, can be part of this component. *E2* expressed his problem with this component as he understood that significant processes involved in data trading have been taken care of by the previous components. In that light, *E2* states that mentioning this service just as a "*download link with SSH*" as a very basic thing to explicitly describe. *E2* remarked that if by data exchange service is interpreted as the network, a connection between 2 end points like saying, "*internet is part of the data market*" which is a very trivial thing in this discussion.

- ***Data Analytics Service:*** *E1* perceived this as ***data analysis service*** being hosted on the *data marketplace platform* or ***a third-party cloud provider***. In that case, *E1* suggested to have a

credential management to verify the legitimacy of these entities before providing access to carry out data analysis on the data. *E2* remarked data analysis services part as very important and suggests 2 variants of provisioning the data analysis services: one variant where *data analysis services are* ***centralised*** and run on the platform and the other variant being the one with *third parties* offering *data analysis services* in an ***app store*** kind of way. Consequently, *E2* was okay with our app store model. However, we change the name of this component to ***Data Analysis Service*** to reflect both the interpretations.

7.1.5.2 Drawing Conclusions

As the focal data marketplace platform of this research was specified to be *many-to-many B2B Decentralised Serendipity Model data marketplace*, the centralised platform design was excluded from the analysis thus narrowing our scope. As a result, the usage of the term (*meta*)*data* to signify both data and metadata being on the platform is no longer used. Furthermore, only *metadata* is managed on the platform centrally while the data resides decentralised. Now, the refinement of the functional components is discussed. The updated list of functional requirements is considered here to newly assign the requirements to the updated conceptualisations of the components.

- ***Identity Management***: The interpretation of this component remains the same with the features of *induction, authentication and authorisation.* However, a new feature is added i.e. ***key management*** as this is a crucial requirement for the materialisation of decentralised data marketplace platform for the enabling and management of encrypted communication channel. Evidently, this component enforces the functional requirement of only *boundary conditions.*

- ***Broker Service:*** This component contains the same 2 features: *Data Management* and *Customer Interaction*. Some of the activities which are part of data management feature remain same while some undergo changes. *Data Cataloguing* and *Data Marketplace Curation* remain the same. *Data Tracking* undergoes a small change with only handling the tracking of *data lineage*. Hence, we rename it as ***Data Lineage Tracking***. Finally, since data does not reside on the platform, the broker service is responsible only for the *management of providing access to the appropriate data* wherever it resides (either on data owner's site or in a distributed system) to the appropriate actors with the help of ***key management***. We term this activity as ***Data Access Management.*** On the other hand, there are no changes in the user interaction service. Coming to the updated requirements, the *broker service* enforces the following functional requirements
  - *Data Provenance* through *data lineage tracking*;
  - *Data Economy* by creating revenue streams for themselves and the actors.
  - *Marketplace Features* though their *data management services*
  - *Platform services* through *user interaction service.*

- ***Clearing House:*** The interpretation of this component also does not undergo any change as it essentially comprises of transaction management system. The component enables *data usage tracking* which involves documenting the usage information of the data like *who has the access to the data, who accessed it and if they accessed or not et cetera.* With this activity, clearing house enforces *data provenance* functional requirement. It can be implemented in different ways. Although the underlying condition is that it should be tamperproof.

- ***Data Exchange Service:*** This component undergoes a major change as a result of the expert insights as our conceptualisation of this component was unclear and very trivial to be a functional component. This component is no longer just a communication channel or a download link with SSH. The data exchange service signifies the ***business process of how the data is shared*** among the involved data actors. In simple words, the logistical way through which the ***data access*** is provided to the data consumer on the data marketplace. The implementation of this component is highly dependent on the use-case and resulting

technical architecture. The concepts like computation, algorithm, data access and even data analysis comes into the picture based on the underlying use-case of data-sharing. With these aspects, the data exchange service enables the functional requirement, *secure data exchange.* Additionally, it goes without saying that in a decentralised design like ours, the data owner has control over his data as he houses the data and access is provided by the data marketplace to the data consumer which is dictated by the use-case of the data sharing. Since this aspect relates to the mechanism of data sharing, this component also enforces the functional requirement of *data sovereignty.*

- ***Data Analysis Service:*** We shall incorporate the additional insight on this component which we got from the experts that *data analysis services* can be also be hosted ***centrally*** on the data marketplace platform. Again, the way to do it is dependent on the business process of the data analysis which is dependent on the use case and the architecture. The feature of app store model still remains with the platform providing data analytic tools from the third parties on the platform in the form of downloadable software or SaaS. The functional requirement of ***data economy*** is satisfied here.

The ***Data Inventory*** component is omitted from our list for a variety of reasons. Firstly, the platform design is decentralised and hence the data does not reside on the data marketplace platform; consequently, eliminating the need for data inventory. Furthermore, in a decentralised setting, where the physical data resides, whether on the client's site or a distributed system or in rare cases in blockchain, does not matter from the perspective of the data marketplace as it is the responsibility of the data owner provisioning the data. The owner provides the access of the data to the broker service which manages that access. These reasons motivated us to remove the component from the list.

In addition to the existing components, we included a new component, ***Governance Model*** to the list of functional components. As discussed earlier in the requirements sections, this component consists of activities which involve enabling the data marketplace platform in the form of trust mechanism, governance or enabling services. These activities are carried out by the *Market Enabling Actors* by designing business processes using technology. Consequently, it fulfils the functional requirements of ***Governance.*** The enabling services can be added as and when necessary according the use-case. Hence, the actors and activities here are not fixed. The following is the updated list of categories and codes associated with the topic, *T4*: *Functional Components of the Data Marketplace Platform* in which the modifications highlighted (additions in ***green*** and deletions in ***red***):

- ***T4:*** *Functional Components of the Data Marketplace Platform*
  - ***C1: Governance Model***
    - ***C1:*** *Identity Management*
      - ***C11:*** *Features*
        - ***C111:*** *Induction*
        - ***C112:*** *Authentication*
        - ***C113:*** *Authorization*
        - ***C114:*** *Key Management*
      - ***C12:*** *Enforced Requirements*
        - ***C121:*** *Boundary Conditions*
    - ***C2:*** *Broker Service*
      - ***C21:*** *Features*
        - ***C211:*** *Data Management Services*
          - ***C2111:*** *Data Cataloguing*
          - ***C2112:*** *Data Marketplace Curation*
            - ***C21121:*** *Data Categorisation*
            - ***C21122:*** *Data Tagging*
        - ***C2113:*** *Data Lineage Tracking*

- ***C2114:*** *Data Access Management*
- ***C212:*** *User Interaction Services*
- ***C22:*** *Enforced Requirements*
  - ***C221:*** *Data Provenance*
    - ***C2211:*** *Data Lineage*
  - ***C222:*** *Data Economy*
    - ***C2221:*** *Revenue Stream*
  - ***C223:*** *Marketplace Platform*
    - ***C2231:*** *Marketplace Features*
    - ***C2232:*** *Platform Features*
- ***C3:*** *Clearing House*
  - ***C31:*** *Features*
    - ***C311:*** *Transaction Repository*
    - ***C312:*** *Data Usage Tracking*
  - ***C32:*** *Enforced Requirements*
    - ***C321:*** *Data Provenance*
      - ***C3211:*** *Data Usage*
- ***C4:*** *Data Exchange Service*
  - ***C41:*** *Features*
    - ***C411:*** *Data Exchange Business Process*
  - ***C42:*** *Enforced Requirements*
    - ***C421:*** *Secure Data Exchange*
    - ***C422:*** *Data Sovereignty*
- ***C5:*** *Data Analysis Service.*
  - ***C51:*** *Features*
    - ***C511:*** *Data Analysis*
    - ***C512:*** *Data Analytics App Store*
  - ***C52:*** *Enforced Requirements*
    - ***C521:*** *Data Economy*
- ***Cx:*** *Data Inventory*

Following the updation of the 3 components of the high-level architecture of the data marketplace from Chapter 2, a new updated high-level architecture is built to reflect the findings obtained so far and represents a more appropriate and comprehensive architecture for a data marketplace platform. The updated architecture is illustrated in Figure 21 which represents the ***Pre-MPC Data Marketplace Platform 2.0***.

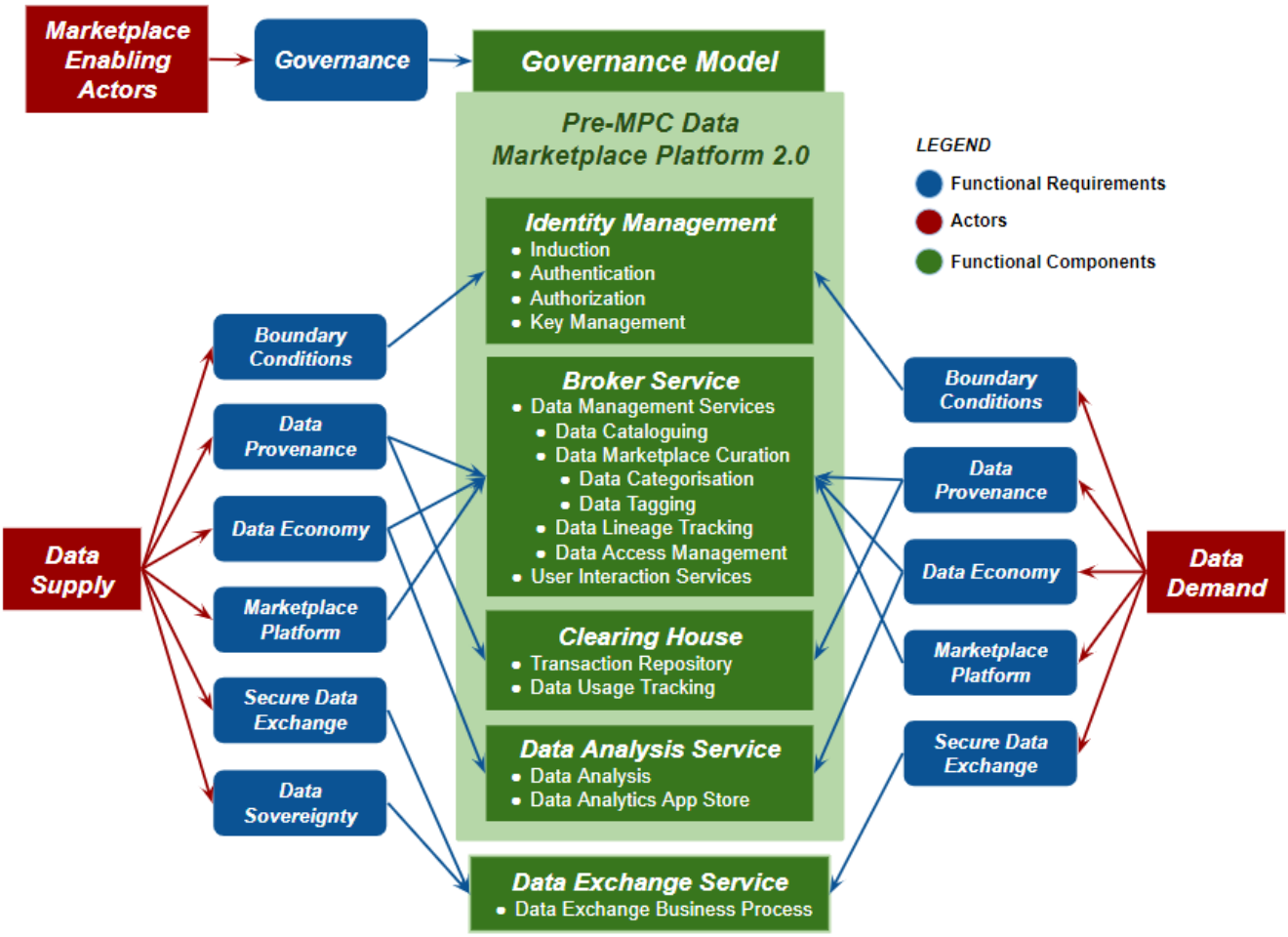


***Figure 21**: Refined High-Level Architecture of the Data Marketplace Platform (Pre-MPC Data Marketplace Platform 2.0)*

Governance Model and *the Data Exchange Service* are intentionally placed outside the platform in Figure 20. The *governance model* comprises of human actors and activities which enforce *governance* on the data marketplace platform by devising various business processes using technology. So, the governance model reflects the coordination between the human actor and technology which collectively enable the functioning of the data marketplace platform. Hence, it did not make sense to include governance model inside the technological architecture of the data marketplace platform. On the other hand, the *data exchange service* is an ad-hoc component which is materialised outside the platform between the data actors involved in the use-case relationship which was established over the platform.

7.1.5 *T5: HLA* Framework

The theoretical concepts of *T6: HLA Framework* from Chapter 2 were not explicitly considered for validation during the expert interviews. However, the updation of the high-level architecture of the data marketplace platform brought about significant changes in the functional requirements, actors and functional components. Hence, it was decided to translate these changes to update the specification of the attributes to obtain an updated HLA framework. The initial list of codes derived from Chapter 2 were,

- ***T5:*** *HLA Framework*
  - ***C1:*** *Functional Requirements*
  - ***C2:*** *Customers*
  - ***C3:*** *Functional Components*

The overall change that the architecture underwent was with respect to its scope. It was understood from expert insights of *E1* that the operations of any technological entity like data marketplace platform cannot be materialised technologically alone but needs a coordinated marriage between human actors and technology. Hence, the scope of the architecture was expanded not only to include the focal technological entity but also the ecosystem that enables the technological entity; basically, the human factor associated with the enabling of the focal technological entity. This change in scope can be propagated to HLA framework as the ecosystem view of the technological entity is more insightful for analysis than the technological one alone. Essentially, the resulting high-level architecture of a technological entity obtained from *HLA framework* will reflect the ecosystem (comprising of human factor) in which the focal entity operates along with its technological architecture. This change brought about changes in all the attributes which reflect the increased scope.

- ***Functional Requirements:*** The modified interpretation of the *Functional Requirements* now reflect not only the technological requirements but also technological ecosystem requirements which reflect the expectations of the actors in the ecosystem from the focal technological entity.
- ***Actors in the Ecosystem:*** The previously termed, *Customer* attribute undergoes major change to expand the horizon to include the human actors along with the customers who enable the focal technological entity. Hence, the attribute was renamed into *actors in the ecosystem*.
- ***Functional Components:*** Similarly, the components comprising of the human activities like auditing, trust enforcement et cetera are also included here now which could give rise to components comprising either solely the human activities or an amalgamation of human and technological activities.

However, the definition of the result architecture would not undergo any change as the it still provides an architecture to a technological entity with surface-level information but not technical specification which applies for either of the technological and human activities. The modified framework is illustrated in Figure 22.

Figure 22*: HLA Framework 2.0*

7.2 *RF2*: Validation of Post-MPC Data Marketplace Platform 1.0

The topics and corresponding theoretical concepts of the research focus, *RF2* are validated here. The artefact under consideration here is the *Post-Data Marketplace Platform 1.0* from Chapter 5. The following 2 topics are validated in the upcoming subsections.

- ***T6***: *Perception of MPC Technology*
- ***T7:*** *MPC Incorporation into the Data Marketplace Platform*

7.2.1 *T6*: Perception of MPC Technology

Since we are not experts in the technical aspects of MPC, our perception of MPC technology is based only on how SafeDEED describes it in their project proposal and the same was understood and incorporated into our study. For this reason, this topic was included here so that our conceptual perception could be validated from the experts and thereby, make the further analysis valid. The validation activity was carried out with the insights obtained primarily from the expert, *E4* who is specialised in MPC technology and works in SafeDEED to implement the technology. *E4* dealt extensively with the value of MPC for data economic market in general. Additionally, experts *E1* and *E2* also provided their insights in this subject which reinforced the insights of *E4*.

7.2.1.1 Results & Analysis

When asked to describe what Multi-Party Computation is, *E4* explains that the basic idea is to bring different parties together to compute something on their inputs without the parties knowing about the inputs of rest of the parties; ultimately learning only the result of the computation and nothing else. But *E4* says that generally this happens with a trusted authority who takes the inputs, computes the function and gives back the result of the computation. Consequently, the authority learns the input data from all the parties. *E4* says that MPC can transforms process into a protocol where the protocol executes the computation, essentially eliminating the trusted authority and still getting the same security guarantee that the result is computed and sent to a dedicated party; without the parties knowing the inputs of the rest of the parties.

E4 relates to the advantages of this property of MPC by mentioning the following. Firstly, the concept of trust is enforced by the system itself and not the actor as there is risk involved. Secondly, *E4* says that with MPC, we can work on data without having to worry about "*leaking personally identifiable information"* in the process. Consequently, *E4* says, "*we wouldn't even need any anonymization techniques because you don't actually have to send the data*" and it is shared through a protocol "*in a randomised way that the others can't learn anything from it*". Adding to this, *E4* further suggests that using MPC, we can carry out computations on private data that is sensitive and that is not legally possible to combine with other data like the "*data from health insurance companies with hospitals as they can't share their databases*". *E4* remarks that the rules around these databases restrict the involved actors to just send their databases to other parties to combine them and compute statistics like "*how often is a person sick? Or are there any other trends like people with higher education get sick less often"*. However, *E4* suggests that MPC allows to compute these statistics because "*the data never leaves your premises in a way that the other party can decrypt it" but* is given access to a protocol that runs the computations and gets only the result. In addition to this, *E4* provides further examples of interesting applications where the property of MPC comes in handy which include "*an auction system where the bis stay private until the final bis is decided*".

Moving on to the logistics of designing and implementing a business process with MPC technology, *E4* states that it starts with a use-case where it makes sense for the companies to interact and share data for which MPC can enforce security for leakage of sensitive private data or confidential

proprietary data that is internal to the companies. *E4* provides an example of a use-case where two companies can combine their customers lists to generate products interesting for the customers in common. Since the list of customers is a confidential proprietary information, they cannot be combined in a traditional way but MPC enables this with one of its protocol called private set intersection (PSI). *E4* stresses that use-case is critical to have beforehand as it will direct the decisions like choosing the protocol, designing the process and running the protocol.

When asked about the 2 MPC processes conceptualised in section 5.2.2, for the interactive process, *E4* confirms that it is a valid process but basic one as different variations of this is possible where everybody receives the output or somebody that is not involved receives the output or some actor only providing computation service over cloud but not providing any data. About the non-interactive process, *E4* disregards the process to be of MPC but rather of traditional computation involving another privacy preserving technology, homomorphic encryption. *E4* reflects that the non-interactive one is a valid process of data aggregation which enables the data owner to provide his data once and not be present every time the computation happens. However, since it is not of MPC technology, the process is out of our scope. Reflecting generally of the processes, *E4* suggests that homomorphic encryption can also be part of the MPC protocol; even data analysis can also be defined as part of the MPC protocol. However, the underlying use-case decides whether the former should be part of the protocol. On this subject, even *E2* reflected confirming that the interactive process is valid representing the true promise of MPC and states that there are many different models of processes which are being developed by his colleagues at *SafeDEED*.

Coming to the limitations of MPC technology, *E4* reflects that the MPC protocol is driven by the function from the use-case. So, it should be made sure "*the function needs to have the property that if you have the input and the function output, then you don't learn anything about the other inputs";* basically, reverse engineering should not be possible with the function. Related to this topic, *E1* also remarks that the application of MPC is currently limited in the real world.

7.2.1.2 Drawing Conclusions

The insights about the basic idea, properties and the advantages of the MPC technology were consistent with what we had dealt. However, the discrepancy with the perception arose in case of processes defined in section 5.2.1. It was presumed that the 2 processes represented 2 kinds of processes of MPC. But it turned out that only interactive process was of MPC and non-interactive wasn't. However, the valuable insight gained in this topic was that of MPC protocol being designed based on an underlying use-case. The use-case being that of data sharing among companies which were suggested earlier by the experts. The fact that the underlying use-case of data computation directs the selection of the function and the design of MPC protocol clarifies that the MPC technology is designed in an ad-hoc way as required by the use-case. This falsifies our perception that MPC is a fixed process like the 2 processes mentioned in section 5.2.1 and that they must be used that way by the actors. On the contrary, the protocol is designed as required by the use-case of the actors. Another useful insight is that the protocol can contain other constituents like homomorphic encryption, different kind of data analysis functions etc. Hence, MPC can carry out many functionalities of data in addition to enable data sharing in a confidentiality-preserving and privacy-preserving way.

However, MPC technology is its own limitations. Firstly, it is still in conceptualisation phase and has not reached maturity as it suffers from scalability issues. Another limitation is that, it is unknown if every function or computation is compatible to be converted into an MPC protocol. The functions derived out of the underlying use-case should be compatible with ***SafeDEED Primitives*** to be converted into a valid protocol. All these limitations should be explored in the future to bring the promises and potential of MPC technology to reality.

7.2.2 *T7*: MPC Incorporation into the Data Marketplace Platform

This topic represents first of the 2 flagship conceptual models contributing towards our research objective as validation of this topic contributes towards the understanding of the architectural implication of MPC technology to the data marketplace platforms. The concepts associated with this topic were analysed by relating them to the insights predominantly of experts *E4* but also, *E1* and *E2*. The contents of section 5.2 drove the list of categories and codes which are listed below.

- ***T7:*** *MPC Incorporation into the Data Marketplace Platform*
  - ***C1:*** *Powers Data Exchange Service*
    - ***C11:*** *SafeDEED Component*
      - ***C111:*** *SafeDEED Primitives*
      - ***C112:*** *SafeDEED Network*
  - ***C2:*** *Enables Decentralised Design*
    - ***C21:*** *Changes Data Inventory to Metadata Inventory*
    - ***C22:*** *Moves Data Sovereignty towards Data Provider's site*
    - ***C23:*** *Moves Data Governance to Data Provider's site*

In the analysis, initially, how MPC technology can be applied generally in data marketplace is discussed and then later, the incorporation is validated for the updated architecture of the data marketplace platform, ***Pre-MPC Data Marketplace Platform 2.0*** from section 7.1.4.

7.2.2.1 Results & Analysis

When asked about what the application of MPC in a data marketplace is, *E4* remarked that the idea of MPC that can work in data marketplaces is that data marketplace can be a platform; where the data owners can say that they have some data and the parties interested in using or running some analysis on that data can connect with the data owner; and then they both can run the MPC protocol privately between them. Evidently, *E4* says that data marketplace can be a place where companies find each other and establish relationship, and the connected companies can install ***SafeDEED Component*** containing the MPC protocol on either of their servers and can carry out data computation. On this subject, *E1* remarks that with MPC, the system itself provides security where the data owners have full control over their data and thereby reducing the need for security governance. *E1* specifically says to enforce decentralised design, MPC makes a huge difference as it eliminates the need for *key management* and the risks associated with it. Sharing this thought, *E2* also says that MPC will play a role in enforcing data sovereignty as the data can no more be misused by the data consumer.

Regarding the changes that MPC technology can bring about in our architecture, *E4* reflects that the components which undergo change with the incorporation of MPC technology would be: ***Data Exchange Service*** and ***Data Analysis Services***. *E4* continued that data exchange service will be transformed with MPC Technology which is enabled by ***SafeDEED Component*** de-centrally running on the connected parties' servers. On the other hand, data analysis service will be moved to the sites of the parties (data owners, data aggregators and data consumers); away from the platform as the data analysis services are run as part of the MPC protocol itself. Other than that, *E4* states that MPC would not affect any other component. *E2* suggests that the data exchange service will be transformed into safer than the traditional way; while also reflecting that none of the other components undergo any change.

7.2.2.2 Drawing Conclusions

Here, we shall reflect what the above findings mean to our research and incorporate the appropriate changes in the updated high-level architecture of our data marketplace platform. The foremost conclusion on the application of MPC technology (foregoing its limitations) is that it

enables a truly decentralised data marketplace platform by truly enabling data sovereignty for the data owners. Furthermore, MPC technology provides *security-by-design* as propositioned in Chapter 5 by truly enabling data sharing and data analysis services in a confidentiality-preserving and privacy-preserving way (where actual data is not known to anybody other than the one who owns it).

The changes brought about with the incorporation of MPC technology into the updated *high-level architecture* from section 7.1.4 are listed as follows.

- The ***Data Exchange Service*** gets transformed from a traditional process (SSH encrypted channel) to a safer and more sophisticated process by including MPC technology through ***SafeDEED component*** (***SafeDEED Primitives*** and ***SafeDEED Network***). The data exchange service will be designed in an ad-hoc way which will implemented in the form of an MPC protocol executing the computation through the ***SafeDEED Component*** running on the servers of all the involved parties.

- The ***Data Analysis Service*** becomes a feature of Data Exchange Service as the data analysis becomes part of the MPC protocol. However, the ***App Store*** component remains on the platform which provides data analytics tools to the actors in the form of downloadable software or SaaS model. So, we shall rename this component as ***Data Analytics AppStore*** to signify its actual meaning.

- The *key management system* in the Identity Management remains but its involvement in the data exchange service depends on MPC protocol if it contains encryption elements.

- Finally, the responsibilities of the security aspect and the trusted authorities are significantly reduced; with the Governance actors not having to worry about the functional requirements of ***Data Sovereignty*** and ***Secure Data Exchange*** as they are fully enforced by the MPC technology.

The updated list of categories and codes representing the effect of MPC technology on the architecture of the data marketplace platform is list below (additions in ***green*** and deletions in ***red***):

- ***T7:*** *MPC Incorporation into the Data Marketplace Platform*
  - ***C1:*** *Enables Decentralised Design*
  - ***C2:*** *Powers Data Exchange Service*
    - ***C21:*** *SafeDEED Component*
      - ***C211:*** *SafeDEED Primitives*
      - ***C212:*** *SafeDEED Network*
    - ***C22:*** *Moves Data Analysis to Data Exchange Service*
    - ***C222:*** *Data Analysis service changes to Data Analytics AppStore*
  - ***C3:*** *Reduces the burden of Governance*
    - ***C31:*** *Enables Data Sovereignty technologically*
    - ***C22:*** *Enables Secure Data Exchange technologically*
    - ***Cxx:*** *Moves Data Governance to Data Provider's site*
    - ***Cxx:*** *Changes Data Inventory to Metadata Inventory*
  - ***C3:*** *Enables Security-by-Design*
    - ***C31:*** *No need for Key Management*

These changes result into the updated *high-level architecture reflecting MPC incorporation*, the ***Post-MPC Data Marketplace Platform 2.0*** as illustrated in Figure 23 (changes highlighted with ***yellow***).

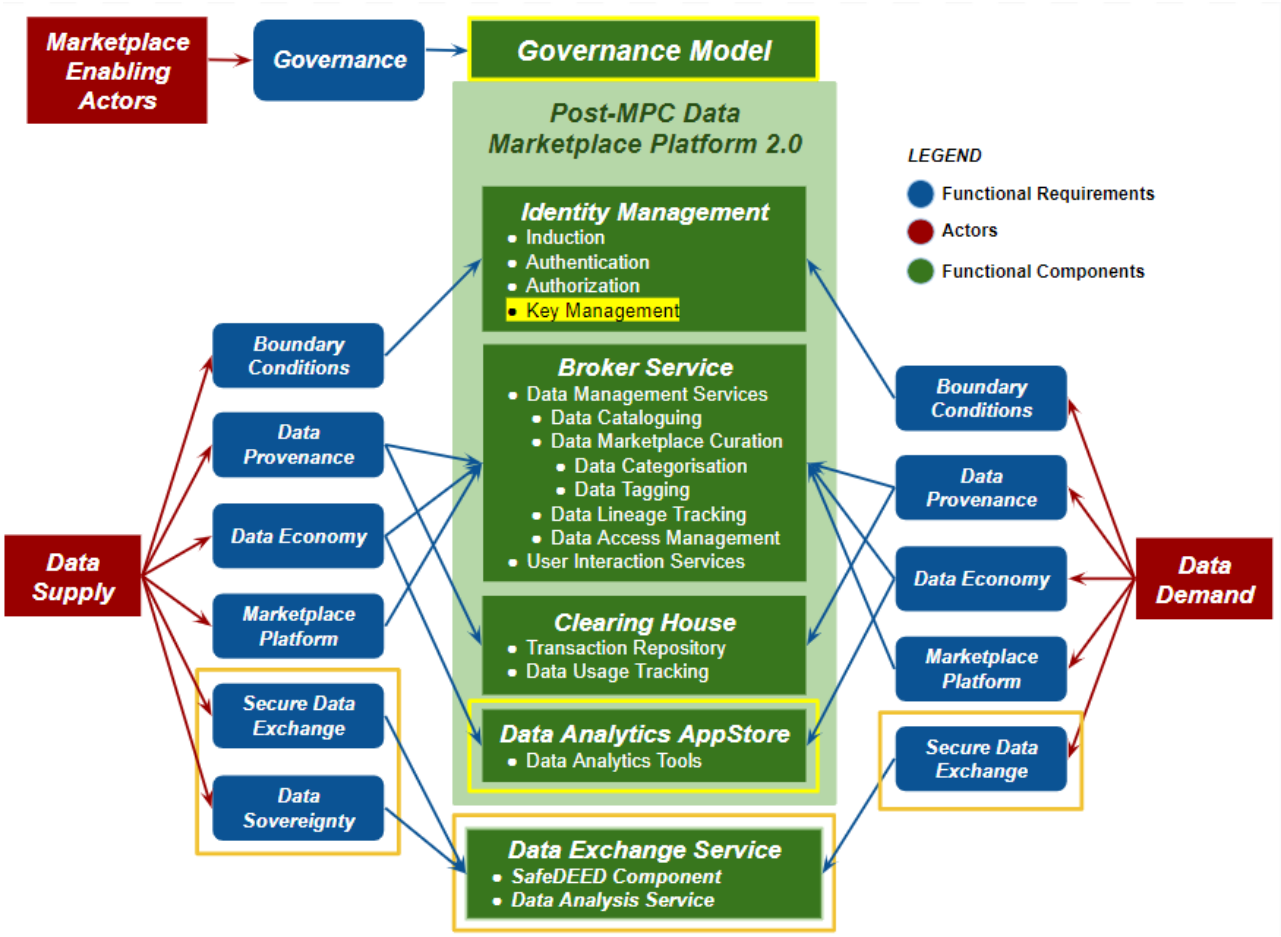


***Figure 23**: Post-MPC Data Marketplace Platform 2.0*

7.3 *RF3*: Validation of Pre-MPC Threat Model 1.0

The topics and their theoretical concepts associated with the research focus, *RF3* are validated here. The conceptual model under consideration here is the *Pre-MPC Threat Model 1.0* built in Chapter 3. This topic was intended to be validated mainly from the cybersecurity expert, *E3* which we did. However, it turned out that expert *E1* also had expertise in this area and *E1* was kind enough to give his insights here. The advantage of having *E1* onboard for this topic was that *E1* is an expert in data marketplaces and hence, we got valid insights related to how to approach the threat aspects of data marketplaces in addition to the process of threat modelling in general. The following 2 topics are validated in the upcoming subsections.

- ***T8:** HLTM Framework*
- ***T9**: Threat Landscape of the Data Marketplaces*

7.3.1 *T8*: HLTM Framework

HLTM framework is a new framework developed by us for the context of high-level threat modelling, and since, threat modelling is a crucial aspect of our research objective, the topic, *T8:*

HLTM Framework was included as part of validation activity. The initial list of categories and codes in this topic derived from Chapter 3 were:

- ***T8:*** *HLTM Framework*
  - ***C1:*** *Context of Threat Modelling*
    - ***C11:*** *Scope*
      - ***C111:*** *At the level of business functions*
    - ***C12:*** *Approach*
      - ***C121:*** *Asset-Centric*
    - ***C13:*** *Purpose*
      - ***C131:*** *Risk Framing – Risk Management*
    - ***C14:*** *Context Statement*
      - ***C141:*** *"to establish the assets associated with the business functions of each functional component of the high-level architecture of a technological entity and later, assume a system specification on which applicable cyberattack vectors (described at a high-level) can be identified"*
  - ***C2:*** *Type of Threat Model*
    - ***C21:*** *High-Level Threat Model*
  - ***C3****: Constructs*
    - ***C31:*** *Functional Component*
    - ***C32:*** *Business Function*
    - ***C33:*** *IT System Asset*
      - ***C331:*** *Primary Asset*
      - ***C332:*** *Supporting Asset*
    - ***C34:*** *Threat*
      - ***C341:*** *Cyber Attack Vector*
      - ***C342:*** *System Failures*
    - ***C35: CIA Violated?***
      - ***C351:*** *Confidentiality*
      - ***C352:*** *Integrity*
      - ***C353:*** *Availability*
    - ***C36:*** *Business Consequence*
    - ***C37:*** *Mitigation Technique*
  - ***C4:*** *Threat Landscape*
    - ***C41:*** *Threat*
    - ***C42:*** *Business Consequence*
  - ***C5:*** *Limitation*
    - ***C51:*** *Baseline Overview*

7.3.1.1 Results & Analysis

When asked about our process of threat modelling using the HTLM framework and the threat model, *E1* reflects that the threat modelling here "*assumes certain implicit architecture*". So, "***the threat model could change if you take a different architectural design***". The *implicit architectural decisions* taken in the *component and business function* construct of how the assets are handled in a data marketplace are an assumption. If the component and business function are implemented architecturally in a different way other than our assumption, then the threat model does not apply. *E1* basically suggests that the threat model will be valid only if there is a defined and detailed underlying architecture. Furthermore, *E1* says that the threat model is valid only to that specific architecture. However, *E1* says that our method is ***fine to obtain baseline threats to the focal entity and hence, its baseline security requirements***. But again, *E1* criticises our threat model to be a "***low-level threat model***" containing threats to a lower level architecture of the components of the data marketplace which will be addressed by the chosen mitigation techniques. But the threats crucial

to the data marketplaces are the ones at the ***higher-level*** like ***"data leakage"*** which are "***difficult to identify***" and "***more complicated***" for our chosen mitigation techniques to prevent; and hence, need "***special mechanisms to mitigate***". *E1* further gave few examples of these higher-level threats to the data marketplaces which will be discussed in subsection xxx when dealing the threats. *E1* suggests that in order to find ***higher-level threats,*** we should understand the main business logic of the data marketplace which is handling data, and hence, we should focus on threats associated with "***data sensitivity***".

E3 reflected overall that the framework and the threat model were relevant and strong compared to the industry standards. However, E3 suggested a few relevant aspects. *E3 suggests that "when looking at the business functions" to do security assessment, we are supposed to consider the* ***processes or procedures****, the* ***requirements*** *towards cybersecurity and how these requirements are enforced within an organization. E3* recommended including ***vulnerabilities*** as a construct as it is the only missing cybersecurity in the framework. About the threat model, *E3* remarked that the threat model is good and comprehensive and suggested few more threats like *system failure, server unavailability*, *malicious insider* et cetera which again belong to the category of "***low-level threats***" of *E1*. *E3* further suggests including threats like ***regulatory, environmental, mismanagement of personally identifiable information*** et cetera to the threat model saying that these are just as relevant as ***IT threats***. Apart from that, *E3* was fine with the framework reflecting that the framework would give ***a generic direction*** towards the security of the focal entity. But *E3* suggested that after generating a high-level threat model, it is necessary to do ***second round of security assessment customised to the specification of the focal entity***. Consequently, *E3* suggested having actual "***architectural concepts of data marketplaces***" in place to "***find valid threats***" echoing the same insight as that of ***E1***. When asked about the significance of high-level threat models generally, ***E3*** echoed our view by saying "*it is a good start to have a set of high-level threats applicable to a type of focal entity*".

7.3.1.2 Drawing Conclusions

The general insight about the HLTM framework and its resulting threat model is that it only represents the starting direction towards the security design of the focal entity. Both *E1* and *E3* reflect this through their "*implicit architecture*" comment (as a result of which it cannot be generalised but only represents a baseline overview) and "*good start to have a set of high-level threats*"; echoing the limitation of our framework that it only provides a baseline security overview which we already have established in the Chapter 3.

Second crucial insight was to go beyond the *IT threats* (*cyber threats*) which is echoed by both the experts. However, our context clearly mentioned the reasoning for this choice that the cyberattack vectors represent the tactic-level description of the technological platform. However, the insight of "***higher-level threats"*** by *E1* is interesting. He basically means that the threats being focussed here are cyber threats operating at the system-level business functions. These threats can be overcome easily through mitigation techniques. However, the threats which exist at a relatively higher-level than the systems' business functions are crucial for the data marketplaces as these threats can disbar the business logic of the data marketplaces are very complex to solve. The example of "data sensitivity" reflects the same that even though the whole system is 100% secure, if an authorised customer behaves malicious where he misuses the data (by leaking it or using for means other than the ones in the contract) which was legally purchased from the data owner. In that case, the mitigation technique could not do anything as everything is working fine but the problem lies in the fundamental business logic of the entity. In the case of data marketplaces, *E1* suggested that the fundamental business logic is the handling of the data and the sensitive nature of protecting it. This relates to the *challenges of commoditising* data that we discussed in Chapter 2. Hence, the nature of data needs to be studied and that knowledge should be applied in the contexts of data marketplaces and gauge how things can go wrong and how that can impact the data marketplaces as organizations. Basically, the data marketplaces should be looked at as business entities than just technological platforms to find the threats that are crucial for the

functioning of the data marketplaces i.e. the threats which actually reflect the threat landscape of the data marketplaces. *E1* further remarks that these threats are "***difficult to identify***" and are "***more complicated***" for our chosen mitigation techniques to prevent; and hence, need "***special mechanisms to mitigate***". We term these threats as ***"business threats"*** as they affect the business logic which is a far higher level than the high-level cyber threats to the business functions of the individual information system'esque components within the entity (which is what performed in HLA framework). This insight was incorporated into the from the NCGI Apex Classification of threat models in the form of a new category of "***business threat models***" which was added at the level beyond the other threat models which only deal only with cyber threats. This is illustrated in Figure 24. Furthermore, in the light of this insight, it was decided to start referring to our focal threats as cyber threats explicitly as they are different from the business threats as mentioned here.

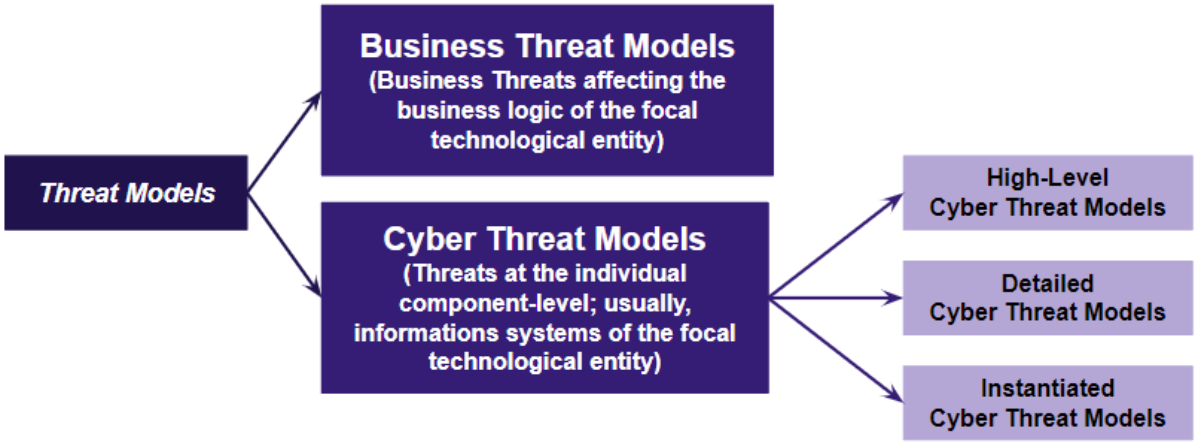


***Figure 24**: Threat Model Taxonomy 2.0*

However, the business consequence construct reflects the effect of the cyber threats to the business functions or the whole focal entity. The latter aspect signifies that the consequence is established not just at the level of information systems or the low-level business functions but al the higher-level of the whole organizations. Hence, this construct reflects the concept of the *business threats* as introduced previously. As a result, the construct can be directly renamed to be called "***business threats***". Furthermore, the insight that these business threats actually represent the threat landscape of the focal entity coincided with our conceptualisation of threat landscape as we had included business consequence as well. It made sense to incorporate threat construct given we were aiming to get a baseline threat landscape. Now that it is established that those threats can be easily overcome by mitigation technologies, it no longer reflects the actual threat landscape of the focal entity. Hence, we shall now refine the conceptualisation to include only the construct of *business threats* alone. However, the cyber threats still contribute here indirectly as they can influence business threats into manifesting. The threat of *"mismanagement of personal identifiable information"* suggested by *E3* reflects this scenario where the breach of PII can affect the business logic of the focal entity is data security was its business logic. However, this scenario applies in the presence of a detailed technical architecture as that guarantees the reflection of actual threat landscape. The refined conceptualisation is illustrated in Figure 25.

***Figure 25**: Conceptualisation of the Threat Landscape 2.0*

Related to the framework, we rename it to ***High-Level Cyber Threat Modelling (HLCTM) framework*** to reflect the cyber aspect revelated earlier. Furthermore, we add the construct of ***vulnerability*** into the framework. Vulnerability is a design flaw that exist in the system under focus which can be exploited by cyberattacks. In the context of the availability of a specific architecture, vulnerability is a relevant construct and hence qualifies to be added into the framework. The resulting cyber threat model will be specifically valid to the architecture under consideration. For high-level cyber threat modelling, however, the construct can be ignored. This move increases the flexibility of the already flexible cyber threat modelling framework. The refined HLCTM framework is illustrated in Figure 26.

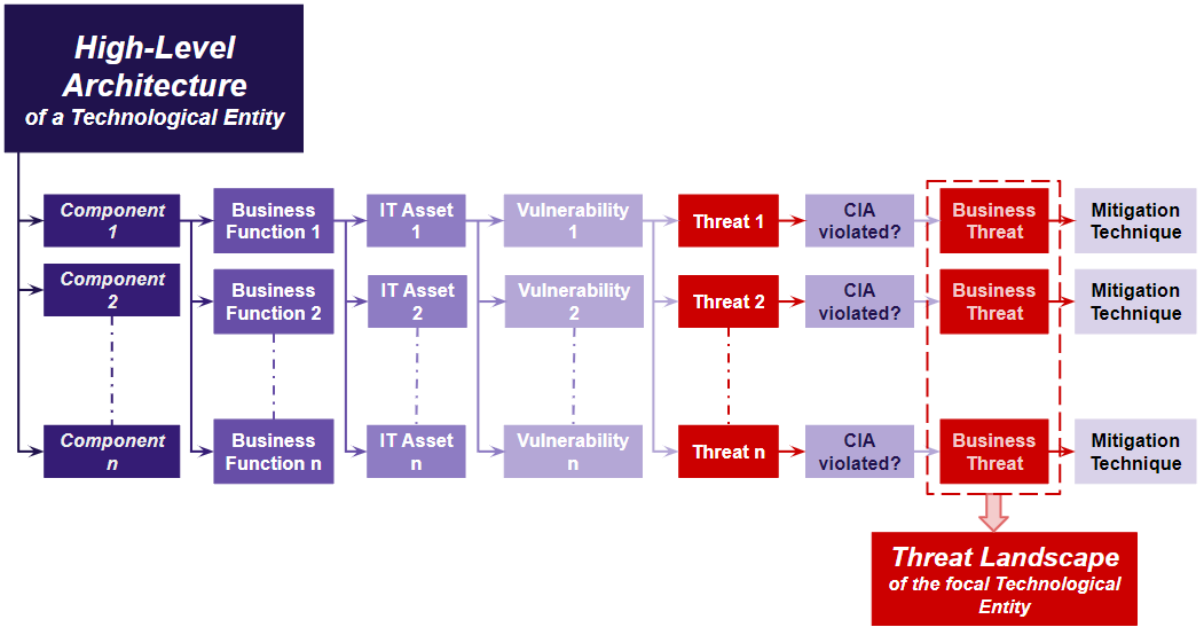


***Figure 26**: High-Level Cyber Threat Modelling (HLCTM) Framework*

The framework was not used again to find the business threats as in the framework, only the cyber threats drive the business threats and are applicable to the detailed technical architectures. However, in our case, this does not apply as our architecture is still high-level. To identify the actual high-level business threats specific to data marketplaces, it is advised either to explore the concept of ***data sensitivity*** and understand the threats around it; or to understand the business logic of a data marketplace by carrying out a case study of a real-life data marketplace and then, identify the business threats to that data marketplace which can then be generalised to all the data marketplaces. However, we used the interviews with the experts to do so which will be discussed in the next subsection.

The updated list of categories and codes associated with *T8: HLCTM Framework* is listed below with the modifications highlighted (additions in *green* and deletions in *red*):

- ***T8:** HLCTM Framework*
  - ***C1:** Context of Cyber Threat Modelling*
    - ***C11:** Scope*
      - ***C111:** At the level of business functions of information systems*
    - ***C12:** Approach*
      - ***C121:** Asset-Centric*
    - ***C13:** Purpose*
      - ***C131:** Risk Framing – Risk Management*
    - ***C14:** Context Statement*

        - ***C141:*** *"to establish the assets associated with the business functions of each functional component of the high-level architecture of a technological entity and later, assume a system specification on which applicable cyberattack vectors (described at a high-level) can be identified"*
- ***C2:*** *Type of Cyber Threat Model*
    - ***C21:*** *High-Level Cyber Threat Model*
- ***C3****: Constructs*
    - ***C31:*** *Functional Component*
    - ***C32:*** *Business Function*
    - ***C33:*** *IT System Asset*
        - ***C331:*** *Primary Asset*
        - ***C332:*** *Supporting Asset*
    - ***C34:*** *Vulnerability*
    - ***C35:*** *Threat*
        - ***C351:*** *Cyber Attack Vector*
        - ***C352:*** *System Failures*
    - ***C36: CIA Violated?***
        - ***C361:*** *Confidentiality*
        - ***C362:*** *Integrity*
        - ***C363:*** *Availability*
    - ***C37:*** *Business Threat*
    - ***C38:*** *Mitigation Technique*
- ***C4:*** *Threat Landscape*
    - ***C41:*** *Business Threat*
- ***C5:*** *Limitation*
    - ***C51:*** *Baseline Overview*

7.3.2 *T9*: Threat Landscape of the Data Marketplaces

Following the discussion from the previous subsection, it was deduced that the *High-Level Threat Model* from the Chapter 4 does not reflect the actual threat landscape of the data marketplace platform but only a baseline overview which does not actually contribute towards understanding the effect of MPC technology on the data marketplace platforms, Hence, although a good and comprehensive cyber threat model, it was discarded to be invalid for our objective. The same was seconded by the expert, *E1* that the threat model does not represent the threat landscape specific to data marketplace platforms. Even though, the threat model had business consequence construct which relates to the actual threat landscape, it was discarded as the value of the business consequence were driven from the baseline cyber threats. Hence, we start from the scratch here with no *Pre-MPC Threat Model.*

However, we got rich and appropriate insights from the experts about the so-called *business threats* which prevail specifically for the business logic of the data marketplace platforms. We gained these insights from the experts *E1* and *E2* who are well versed in the field of data marketplaces and hence, reflected well on the subject of the threats associated with them. In addition to this, *E4* also contributed with a threat scenario which applies here. We present the same insights and analyse them to generate the valid list of threats which do reflect the actual threat landscape of the data marketplace platforms, reflecting the *Pre-MPC Threat Model 2.0*. Furthermore, the new list of threats is generated by conducting ***Open Coding*** through which codes are generated form the data without any initial list of categories and codes. As a result, we end up with a fresh list of categories and codes straight from the data.

7.3.2.1 Results & Analysis

With respect to the threats associated with the data marketplace platforms, the experts talked mostly about high-level business threats while giving threat scenario examples for few threats. In this subsection, we shall present the threats in a qualitative way as described by the experts and codify accordingly.

E1 tells that the issues that are currently crucial to the data marketplaces are not the attacks from external entities; but the internal problems within the data marketplace. According to *E1*, these are the issues being worked and researched on by the industry rather than the cyberattacks. *E1* gives examples of these issues and they are quoted here with their respective labels that we have coded.

- *"once the consumer gets the data, nothing stops the consumer to do whatever he wants with the data".* We reduce this statement into 2 codes, ***Loss of Control over Data*** and ***Data Leakage;*** and assign it to the category of ***Threats.*** Furthermore, relating about mitigating, *E1* says, *"this comes back to governance".* We code this with label, ***Governance*** and assign it to the category of ***Mitigation Techniques.***

- *"the internal actors have to work solely on the trust over the parties".* We code this with label, **Trust Issues** and assign it to the category of ***Threats.***

- In decentralised design, where the access of data is given over an encrypted channel, on the receiving end, *E1* states that the algorithm at the receiving end can copy the data by saying, "*the container with an algorithm which need to decrypt the data. However, the container is going to copy the data somewhere. So, the basic security is broken. You have to do much to work to enforce which involves governance*". Evidently, the ***threat*** aspect of this statement can be coded into ***Loss of Control over Data*** and ***Data Leakage***; and the ***mitigation technique*** to be ***Governance***.

- "*the data providers are going to trust you with his credentials, and some would not trust and would not give the keys*"**. *Threats > Trust Issues*.**

- "*You have higher level threat models like of data leakage. For example, you're sharing data under some contract which restricts its usage and so on. The threat models where the data can be correlated with another data set which can lead to some leakage. For example, you try to anonymize the data, for example, by removing some items like personally identifiable information. One of the threat models is that this anonymized data can be correlated back to the identities if it is combined with other appropriate datasets. So, these are the high level threat model cases that still need to be addressed in the setting of a data marketplace."* We code this account with the label, ***Data Leakage by Back Correlation*** and categorise it to the ***Threats***.

- *"are high-level like data leaks. For example, if you provide a database of all the people in the Netherlands, and you try to anonymize it, and you say you can use this data, then there exists a threat model that someone would get this anonymized database and correlate it to the identities. The threat model referred runs as an application which corelates the anonymized data to the identities."* Same case as the previous account. Hence, ***Threats > Data Leakage by Back Correlation.***

- When talking about how to identify these high-level business threats, "*I came across them mostly from talking about data sensitivity*"**.** Hence, we code this statement as ***Data Sensitivity*** and categorise as broad one in ***Threats.***

- *"And with a lot of AI being done now, there's all this back correlation of census data".* This is another case of the same code, ***Data Leakage by Back Correlation***.

- *"The problem with sensitive data is that it can relate logically. So architecturally, you have everything secure. But the algorithm that's being applied on the data can itself be a threat as it can cause data leak. i.e. if the data is not properly anonymised, then the leaked data itself*

can be sensitive even without back correlation. So, in this example, even how to anonymize can be a big issue." This is a simple case of data leakage and hence the code label is ***Data Leakage***. The statement can also be coded to ***Data Sensitivity***.

- "*for example, the MRI images. You can say the MRI images itself can be processed and with tracking preferences, determine that the MRI image itself is already identifying people because it becomes like a fingerprint. Although it's anonymized and it doesn't mean anything to get the MRI images, but by correlating it with image processing and other dataset, the sensitive data can find the home it belongs."* This is again an example for a case of the same codes, ***Data Leakage by Back Correlation*** and ***Data Sensitivity.***

- "*I don't think there's a way to mitigate these risks to hundred percent. That's where you come back to governance issue because eventually, when there is some data leak, you have the auditors and everything that you can litigate legally."* This statement clearly belongs to the code, ***Governance*** of the category ***Mitigation Techniques.***

Now, moving on to our next expert, *E2* provided some additional inputs towards the subject of threats associated with the data marketplaces. *E2*'s insights are quoted below and are coded and categorised accordingly.

- *"As long as somebody has access to the data, they can write a function on it, that copies the data, and then, subsequently misuse it. That is their intention".* This statement can be related to the code of ***Loss of Control over Data*** and ***Data Leakage;*** and categorise it to ***Threats***.

- "they will lose control over their data and that the data will be out in the wild, even if it is behind the paywall. Somebody else will pay for it and then they will release it. They give the example of obviously, movies or whatever. They are all behind the paywall, and then somehow, they all ended up on some BitTorrent site". ***Threats > Loss of Control over Data*** and ***Threats > Data Leakage***

- "*in production and manufacturing, producing data from the machines has the potential danger of a competitor reverse engineering their processes. For instance, it can be like, they have a special process that they produce some plastic at a certain temperature, which makes it better or more stable. And then if they release sensor data from the machines about energy consumption and operation times, then based on the energy consumption, perhaps the competitor will be able to determine the temperature they're using in the process. This is an example in the industry and manufacturing".* Classically, this instance appends to the code of ***Data Leakage.*** However, this account brings about a new code with label, ***Loss of Competitive Advantage for Data Actors*** caused by the disclosure of proprietary information. This is categorised to ***Threats.***

- *"In all other sectors like Banking, Telecom or the Health, of course, the problem is with the regulations. They are afraid at some point that the data will be deanonymized and therefore, they will be facing fines for having released personally identifiable information."* This statement can be coded with label, ***Regulatory Threats*** and ***Data Leakage by Back Correlation*** (deanonymization is done through back correlation of data). Additionally, it is categorised into ***Threats.***

In addition to these, *E4* suggested a threat scenario which applies in this situation. This is shown in a qualitative way below and is coded accordingly.

- *"if you are a malicious actor, and you use the marketplace, and you do computations with everyone; but you always just make up all the data, then it doesn't look too good from the perspective of the marketplace".* This statement is reduced to a new code, ***Induction of Malicious Data Actor*** and is categorised under ***Threats***

Ultimately, under the topic of Threat landscape of the Data Marketplaces, we shall have 2 categories: ***Threats*** and ***Mitigations***. In the category of *Threats*, we end up the codes *Loss of Control over Data, Trust Issues, Data Leakage*, *Data Leakage due to back correlation of data, Loss of Competitive Advantage, Regulatory Threats, Data Sensitivity* and *Induction of Malicious Data Actor*. On the other side, the category of *Mitigation Techniques* consists of only one code, *Governance.* The final codes and categories are listed in the Table 22.

***Table 22**: Updated Categories and Codes and their number of references by Experts*

Category	Code	Suggested By	No. of Instances mentioned	Total No. of instances
Threats	Loss of Control over Data	E1	2	4
		E2	2	
	Trust Issues	E1	2	2
	Data Leakage	E1	3	6
		E2	3	
	Data Leakage by Back Correlation	E1	4	5
		E2	1	
	Loss of Competitive Advantage for Data Actors	E2	1	1
	Regulatory Threats	E2	1	1
	Data Sensitivity	E1	3	3
	Induction of Malicious Data Actor	E4	1	1
Mitigation Techniques	Governance	E1	2	2

7.3.2.2 Drawing Conclusions

Here, we shall convert the codes and categories obtained from the previous subsection into the Business Threat Model in lines with the objective of this research. The Business Threat Model is illustrated in the Table 23. The constructs used here in this threat model are: Business Threat, Threat Description, Threat Experiencing Actor and Mitigation Technique. The threats are described as appropriate to the *Pre-MPC Data Marketplace Platform 2.0*. However, different interpretations of these threats apply in all the designs of the data marketplace platforms. The Business Threat Model, also representing ***Pre-MPC Threat Model 2.0***. represents the ***actual threat landscape*** of the data marketplace platform.

***Table 23**: Pre-MPC Threat Model 2.0*

Business Threat	*Threat Description*	*Threat Experiencing Actor*	*Mitigation Technique*
Loss of Control over Data	The threat comprises of instances where once the data is transacted and is away from the data owner, the data can be exploited to do anything. It can be used for malicious activities, or it can be resold to some other actor or it can simply be copied and released over internet. Since the data owner legally owns the data and licenses it to the consumer, then if that data is used by the consumer for malicious activities, then even the data owner will be held legally liable for that malicious act since the data he legally owns was used there.	Data Owner	***Governance Model***
Trust Issues	Since the actors are expected to participate based only on the trust towards the marketplace authority and the other data actors and since, there is no tangible way of proving the trust mechanism in place and also unavailability of any technological way of enforcing trust, the data actors may not participate in the data marketplaces as they don't trust somebody else with their data. This turns out to be a threat to the Marketplace provider.	Marketplace Provider	
Data Leakage	This threat is a straight forward one where the data being transacted gets used by the involved data actors in a way that was not intended by the data owner in the terms of the contract. So, the data is being used as not intended which is a threat to the data owner and also to the supply side as they are also involved in processing the data to be transacted.	Supply Side Actors	
Data Leakage by Back Correlation	A special kind of Data Leakage threat where the data with personally identifiable information (PII) is anonymised and transacted to the consumer; and then, the PII can be extracted from the data either because of faulty anonymization or by combining it with other auxiliary data sets, and eventually correlating it back to the original PII.	Supply Side Actors	
Loss of Competitive Advantage for Data Actors	This is a different kind of threat resulting out of data leakage threat where, from the shared data, the receiving actor learns some proprietary information about the data owner selling the data or the supply side actors involved in the business process of the data transaction. The case of back correlation or combining with other data by the receiving actor can result in the loss of competitive advantage to any of the applicable supply side actors.	Supply Side Actors	
Regulatory Threats	This is the legal aspect of all the threats covered here. The threats discussed till now can result in regulatory threats for various reasons like violation of the terms in the contract, violation of privacy et cetera. The logistics of how exactly this threat apply is dependent on specific cases.	All Actors	

Data Sensitivity	This is a broader threat which relates to the unique characteristics of data that makes it challenging to commoditise it as discussed earlier in section 6.2.1. It can be stated that all the threats associated with data are due to this broad threat of Data Sensitivity. All the threats dealt before this here can be stated as specific cases resulted because of the sensitive nature of the data.	All Actors	
Induction of Malicious Data Actor	This is a generic threat to any marketplace where a malicious actor is inducted into the data marketplace as a legitimate costumer. The data actor can be on the platform to exploit the services which is a risk to the data marketplace. This data actor can provide bad data for the computations, thereby generating invalid results for the fellow actors.	All Actors	

The final list of the codes in the topic *T9* are as follows:

- ***T9: Threat Landscape of the Data Marketplaces***
  - ***C1:*** *Loss of Control over Data*
  - ***C2:*** *Trust Issues*
  - ***C3:*** *Data Leakage*
  - ***C4:*** *Data Leakage by Back Correlation*
  - ***C5:*** *Loss of Competitive Advantage for Data Actors*
  - ***C6:*** *Regulatory Threats*
  - ***C7:*** *Data Sensitivity*
  - ***C8:*** *Induction of Malicious Data Actor*

Coming to the Mitigation Technique aspect of the Business Threat Model, the experts feel that the technology is not mature enough to address the issue of ***Data Sensitivity*** and its ramifications (other threats resulting out of ***data sensitivity***); and hence, cannot enforce the functionalities of the data marketplace platform in a comprehensive way. They still feel that it is a collective coordination between the technology, regulation and the actors involved complementing into a ***Governance model*** which enforces every functional requirement along with the security. However, it goes with saying that a 100% security is never possible, and the threats are never mitigated or eliminated but only minimised. The effective enforcement of the ***Governance Model*** is the solution towards minimizing the threat influence on the data marketplaces.

7.4 *RF4*: Validation of Post-MPC Threat Model 1.0

The topics and corresponding theoretical concepts of the research focus, *RF4* are validated here. The artefact under consideration here is the *Post-Data Marketplace Platform 1.0* from Chapter 5. The intention here is to investigate the effect of MPC on the threats associated with the data marketplaces. Consequently, the topic validated here is,

- ***T10***: *Effect of MPC Incorporation on the Threat Landscape of the Data Marketplace Platform*

7.4.1 *T10*: Effect of MPC Incorporation on the Threat Landscape

This topic represents second of the 2 flagship conceptual models contributing towards our research objective as validation of this topic contributes towards the understanding of the implication of MPC technology to the threat landscape of the data marketplaces. As the high-level threat model from Chapter 4 was deduced to be invalid to our objective as it does not reflect the threat landscape of the data marketplaces, the same thing applies to the *Post-MPC Threat Model 1.0*. As a result, there is not validation in this section. But the final conceptual model of this research ***Post-MPC Threat Model 2.0*** is developed with the insights of experts, *E1* and *E4*.

7.4.1.1 Results & Analysis

E1 suggests that even after MPC is in place, the threat of trust issues exist in the sense that since the governance authority is eliminated and the trust surrounding data is totally handled by MPC technology, the data owner might find it difficult to trust the other data actors in the absence of the governance authority. Hence, *E1* recommends some form of governance to tackle these threats. This issue relates to the business threat of **trust issues** in the Pre-MPC threat model 2.0.

On this subject, *E4* remarks that with MPC in place and the absence of governance authority, the threat of malicious data actor increases as the trust mechanism is maintained by technology and the malicious data actor will get away with providing faulty data and using the service and resources of the data marketplace. However, the threat of malicious data actor impacts severely on the other data actors while only causing only reputation loss to the marketplace provider. But *E4* reflects that this may escalate if there are more malicious data actors than legitimate and honest data actors. In that case, *E4* reflects if contracts between the data actor and the marketplace provider are set up and hold accountable legally if the data actors behave maliciously. This relates to business threat of ***malicious data actor*** in the Pre-MPC threat model 2.0.

7.4.1.2 Drawing Conclusions

The Post-Threat Model 2.0 comprising of the business threats which prevail for the data marketplaces even after the incorporation of MPC technology is generated here. Firstly, the effect of MPC technology on each business threat in the ***Pre-MPC Threat Model 2.0*** is discussed and the ones which will be mitigated by MPC are filtered. Parallelly, the business threats post-MPC as identified from the insights of the experts to finally obtain the ***Post-MPC Threat Model 2.0.***

- ***Loss of Control over Data:*** Since there is no actual transfer of the data from the data owner to the other actors and the fact that the data resides at the site of the data owner and provisioned remotely through MPC protocol, the business threat of loss of control over data does not apply anymore. However, the terms of how the data is provisioned for the MPC protocol by the data owner should be stipulated over the contract and there should be governance model to enforce this.
- ***Trust Issues:*** This business threat transforms into a different case of trust issue where the data actors find it difficult to trust other data actors in a technological setting with the ***absence of the authority***. To tackle this, since MPC technology ensures the enforcement of trust surrounding the data, there should be a ***governance model*** to handle the trust associated with the rest of the aspects of the data marketplaces.
- ***Data Leakage:*** On the assumption that MPC protocol works efficiently and effectively, since there is no actual transfer of the data between the actors, the business threat of data leakage does not apply any more.
- ***Data Leakage by Back Correlation:*** Here, the same thing applies as the previous threat and hence, even this business threat is no more applicable.

- ***Loss of Competitive Advantage for Data Actors:*** The same reason as the previous 2 business threats apply here too. However, this can depend on the function or the data analysis service in the MPC protocol as the receiving actor can further analyse the computation result by combining it with other auxiliary data or reverse engineering et cetera. With the effective and efficient execution of the MPC protocol, this business threat does not apply.
- ***Regulatory Threats:*** The same reasoning as the previous business threat applies here where in the intended functioning of MPC protocol, the business threat does not apply. However, the business threat can apply in extreme cases of data leakage due to faulty execution of MPC protocol.
- ***Data Sensitivity:*** This is the business threat that MPC Protocol is specifically tackling, addressing and mitigating. Since MPC protocol ensure the functional requirements associated with data sensitivity, ***Data Sovereignty*** and ***Secure Data Exchange***, the business threat of data sensitivity does not apply anymore.
- ***Induction of Malicious Data Actor:*** With MPC protocol in place and the absence of governance model, the business threat of malicious data actor increases as the trust mechanism is maintained only by the technology. The business threat can escalate to a detrimental level when the number of malicious data actors present on the platform exceed the number of legitimate and honest data actors. Hence, this business threat prevails as it affects the functional components of ***Boundary Conditions*** and ***Secure Data Exchange*** and can be addressed through a ***stricter induction*** of data actors as part of ***governance model*** to enforce non-data sensitivity related trust governance while letting MPC technology to enforce data sensitivity related trust maintenance. Furthermore, a more sophisticated contract management enforcing the terms of the contracts between the data actors and marketplace provider can be incorporated. Perhaps, upgrade the contract management with Blockchain Technology.

This brings us to the end of our discussion about the effect of MPC technology on the threat landscape of the data marketplaces. During this discussion, we came across the issue of the effective and efficient execution of the MPC protocol where if this is compromised, all the business threats mitigated by the MPC technology shall return and apply again. Hence, we shall include this as a business threat of ***Faulty Execution of MPC protocol*** in the threat model. With this business threat, the uncertainty involved with the business threats of loss of competitive advantage and regulatory threats is addressed. *Faulty execution of MPC Protocol* can be mitigated by employing auditing authority who can carry out auditing of the MPC protocol and its associated processes, essentially qualifying to the mitigation technique of ***Governance Model*** which includes ***MPC process auditing***. The resulting threat model is the ***Post-MPC Threat Model 2.0*** and is shown in the Table 24.

***Table 24**: Post-MPC Threat Model 2.0*

Business Threat	*Threat Description*	*Threat Experiencing Actor*	*Mitigation Technique*
Trust Issues	The threat of data actors not participating in the data marketplace as the data actors find it difficult to trust a technological setting of just MPC protocol to handle their valuable commercial data	Marketplace provider	***Governance Model with MPC Process Auditing***
Induction of Malicious Data Actor	The threat of inducting a malicious actor into the data marketplace as a legitimate costumer. With MPC protocol in place and the absence of governance model, the threat of malicious data actor increases as the trust mechanism is maintained only by the technology. The threat can escalate to a detrimental level when the number of malicious data actors present on the platform exceed the number of legitimate and honest data actors; affecting the functional components of ***Boundary Conditions*** and ***Secure Data Exchange***	All actors	
Faulty Execution of MPC Protocol	The compromise of the intended (effective and efficient) execution of MPC protocol which can result in the return of all the threats associated with data sensitivity.	All actors	

In the ***Post-MPC Threat Model 2.0,*** it can be seen that MPC technology eliminates the business threats of ***Loss of Control over Data, Data Leakage, Data Leakage by Back Correlation, Loss of Competitive Advantage for Data Actors, Regulatory Threats*** and ***Data Sensitivity.*** So, basically, MPC eliminates the business threats associated with the issue of ***Data Sensitivity*** and its ramifications; and MPC does so still in a ***Security-by-Design*** way as mentioned in Chapter 5; thus, reducing the burden of the ***Governance Model*** on its technological front.

This marks the end of our *Validation phase*. The resulting updated conceptual models from this phase are referred as ***Artefacts 2.0*** and are list below,

- *Pre-MPC Data Marketplace Platform 2.0*
- *Post-MPC Data Marketplace Platform 2.0*
- *Pre-MPC Threat Model 2.0*
- *Post-MPC Threat Model* 2.0

These updated conceptual models and the corresponding updated theoretical concepts are used to answer the sub-research questions associated with the *Validation Phase.* The answers are discussed in chapter 8, where we answer all the sub-research questions along with the main research question, *RQ* of the thesis.

8

Conclusions and Discussion

The research in this thesis was conducted to contribute towards the broader problem of realising data marketplaces into achieving mainstream adoption by data actors. On a narrow note, the prospect of the maturation of Multi-Party Computation (MPC) technology, proposed by SafeDEED**:** *Safe Data Enabled Economic Development*, for the realisation of data marketplaces was the underlying focus of the research. This was performed for the 2 pressing issues, *architecture* and *threat landscape* which were deduced to be the most relevant barriers for the realisation of data marketplaces. On these lines, the research objective of the thesis was formulated to be,

RO: "To understand the implication of the maturation of Multi-Party Computation (MPC) technology for the architecture and the threat landscape of the Data Marketplaces"

The research objective was translated into an exploratory question that served as the main research question of the thesis, which was formalised to be,

RQ: What can be the implication of the maturation of Multi-Party Computation (MPC) technology for the architecture and the threat landscape of the Data Marketplaces?

To answer this question, *4 conceptual models* were developed initially by desk research and then, were validated through *expert interviews* followed by *qualitative data analysis* on the lines of *Middle-Ground Approach* of theory generation. The resulting conceptual models which are validated, contribute towards the implication of the maturation of MPC technology to the architecture and the threat landscape of the data marketplaces respectively are,

- Architectural Implication of MPC technology to the Data Marketplaces
  - *Pre-MPC Data Marketplace Platform 2.0* (pre-MPC architecture)
  - *Post-MPC Data Marketplace Platform 2.0* (post-MPC architecture)
- Implication of MPC technology to the Threat Landscape of the Data Marketplaces
  - *Pre-MPC Threat Model 2.0* (pre-MPC threat landscape)
  - *Post-MPC Threat Model 2.0* (post-MPC threat landscape)

The answer to the main research question signifies a theoretical framework whose conceptual models and corresponding hypotheses are presented in this chapter. With the help of these deliverables, this thesis contributes towards the ongoing research of data marketplaces and the business application of MPC technology.

The rest of the chapter is structured as follows. Section 8.1 presents the theoretical framework with 2 conceptual models and corresponding hypotheses which reflects the main deliverable of the thesis. Section 8.2 provides the answers to all the sub-research questions which were formulated and answered during the course of the research. Section 8.3 discusses the specific theoretical and practical contributions of the thesis. Section 8.4 reflects on the limitations suffered by the thesis. Section 8.5 recommends future work directions in the realm of the focal research problem of the thesis. Finally, section 8.6 concludes the chapter touching upon the relevance of the thesis towards the program of Management of Technology at the TPM faculty of the Delft University of Technology.

8.1 Resulting Theoretical Framework & Conceptual Models

Firstly, the 2 conceptual models are built by formulating their subsequent hypotheses for the implication of the MPC technology to the architecture and the threat landscape of the data marketplaces respectively. Following these, the theoretical framework as per the research objective is developed which answers the main research question, *RQ*. All these are presented in the following subsections.

8.1.1 Architectural Implication of MPC to the Data Marketplaces

The conceptual model representing the first half of the theoretical framework is described here. The conceptual model reflects the implication of the MPC technology to the architectural aspects of the data marketplaces by explicating the difference between *Pre-MPC Data Marketplace Platform 2.0* and *Post-MPC Data Marketplace Platform 2.0*; further generalising the same for the business species of data marketplaces. The hypotheses formulated as a result are listed as follows and the corresponding conceptual model is illustrated in the Figure xx.

Enables Data Trading in a Confidentiality-Preserving and Privacy-Preserving way

MPC technology could enable data trading and data sharing to happen in a confidentiality-preserving and privacy-preserving way where the data owners do not have to transfer the physical data to the receiver. Instead, the data sharing process is converted into a *cryptographic protocol* through which only the result of the computation on the data (or the union of data in case of multiple parties) is shared with the dedicated receiver(s) with the actual input data not revealed to any of the parties involved in the transaction. Since the transfer of physical data is not present, MPC improves the business potential of data trading for all the actors involved in the data marketplace ecosystem. Additionally, the need of anonymization for data owners becomes irrelevant because of privacy-preserving nature of the MPC protocols.

Transforms the Data Exchange Service to ensure Secure Data Exchange

The traditional *Data Exchange Service* which was assumed to an SSH encryption-based communication channel, associated with a vulnerable and costly *key management system* which involves physical data being encrypted and sent over the channel to the receiver who can obtain the decrypted form of the physical data. If the receiver uses the received data for supposes other than the terms of the transaction, there is no way of knowing that because of the *non-rivalrous* nature of data (that it can be replicated at negligible cost and used simultaneously). With MPC technology, it could be transformed into a safer and sophisticated MPC protocol which not only

eliminates the *key management system* for the data marketplaces but also enforces the functional requirement of *Secure Data Exchange* for the data marketplace ecosystem.

Enables Data Sovereignty for the Data Marketplace to be truly-decentralised

Because of MPC technology, the physical data could no longer be transferred to other entities. Instead, the data owners could hold the data with themselves and could provision it to the dedicated receivers like data aggregators or data consumers et cetera with the MPC protocol which eliminates the need for a governance authority to overlook this process. Through this, the MPC protocol enforces the functional requirement of *Data Sovereignty* for the data owners thereby overcoming the challenges of commoditising data: *Protection Regime* and *Quality Control*. This property enables the data marketplace to be *truly decentralised* which makes it a trustworthy platform for data trading.

Supports Data Analysis Services

MPC protocols enable data analysis service as part of their protocols; thus, could lighten the data marketplace platform of the responsibility and infrastructure of data analysis services. It also enables the data analysis service providers to operate trust-free, independent of the data marketplace platform.

Reduces the burden on Governance Model

Since MPC technology alone enforces the functional requirements of *Data Sovereignty* and *Secure Data Exchange* technologically, its incorporation eases the responsibilities of the *Marketplace Enabling actors* with respect to *Data Governance.* However, they still must look after *Marketplace Governance.*

The corresponding conceptual model representing the implication of architectural aspects to the threat landscape of the data marketplaces is illustrated in the Figure 27.

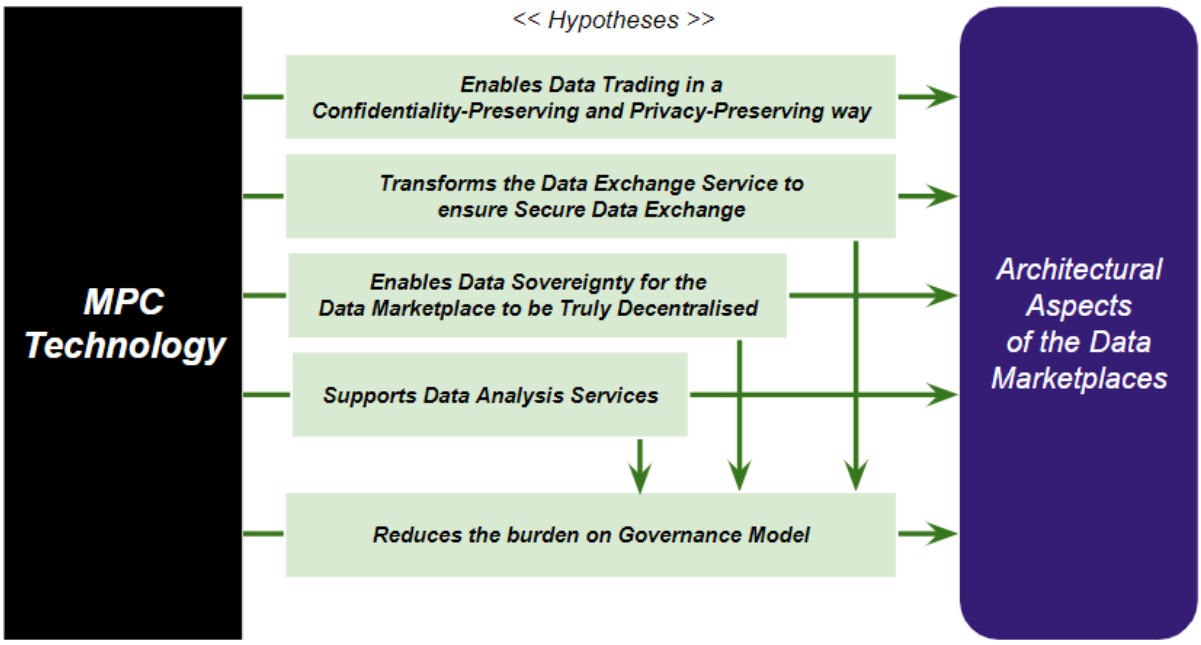


Figure 27: *Architectural Implication of MPC technology to the Data Marketplaces*

8.1.2 Implication to the Threat Landscape of Data Marketplaces

The conceptual model representing the second half of the theoretical framework is described here. The conceptual model reflects the implication of the MPC technology to the threat landscape of the data marketplaces by explicating the difference between *Pre-MPC Threat Model 2.0* and *Post-MPC Threat Model 2.0*; further generalising the same for the business species of data marketplaces. The hypotheses formulated as a result are listed as follows.

Affects the Business Threat Landscape both positively and negatively

MPC technology could mitigate few of the business threats associated with the sensitive nature of data. This aspect could be attractive for all the actors in the data marketplace ecosystem as it overcomes the main concerns associated with commoditising and trading data. On the other hand, MPC could also affect negatively on some of the threats posing more threat than mitigating them. Mitigation of each business threat identified earlier is described as follows,

Mitigates the threat of Loss of Control over Data

MPC incorporation could enforce *Data Sovereignty* effectively enabling the truly decentralised data trading platform which overcomes the threat of data owners losing the control over their data as they hold their data at their site and MPC protocol provides the required knowledge from the data to the dedicated receiver with the help of its cryptographic blocks.

Decapacitates Data Breach

MPC protocol holds either encrypted version of the data or the intermediate data in during the protocol execution which decapacitates the threat of data breach on the communication channel as the breached data does not have any value. This way the threats of *Data Leakage* and *Data Leakage by Back Correlation* could become irrelevant by MPC incorporation. However, the breach could disrupt the protocol execution. However, the risk associated with this could be very less compared to the actual data breach.

Ensures no Loss of Competitive Advantage for Data Actors

As the data exchange happens in a confidentiality-preserving way via MPC Protocol; in the sense that only the result of the agreed upon computation is learnt to the receiver, there would be no risk of that receiver reverse engineering critical aspects of the owner's business processes; thus, MPC could overcome the threat for the data actors losing their competitive advantage when sharing data.

Decapacitates Regulatory Threats because of Privacy-Preservation

The privacy-preserving nature of the MPC Protocol preserves the personal information which could be in the data provisioned by the data owner. This is an incentive for the data owners as the regulatory threats associated with privacy violation and data security are made irrelevant because of no physical data transfer or access.

On the flip side, MPC incorporation could present with shortcomings to the existing situation. The business threats that apply for the data marketplaces even after the incorporation of MPC technology into the business processes of the data marketplaces can be attributed as the negative

implication of the MPC technology to the threat landscape of the data marketplaces. These are hypothesised as follows.

Redefines the threat of Trust Issues

The threat of *Trust Issues* could get redefined into data actors not wanting to participate on the data marketplace platform as they could find it difficult to trust a technological setting of just MPC protocol to handle their valuable commercial data.

Intensifies the threat of Induction of Malicious Data Actor

MPC could incentivise malicious data actors who just wants to gain from the benefits of the data marketplace platform either by making relationships to gain insights, supplying faulty data for the protocol execution, gaining intelligence of other actors who might be competitors et cetera. If the number of malicious parties in the execution of protocol exceeds the number of honest parties, then the protocol would become invalid. This can again be attributed as a ramification of the threat of trust issues where in the technological setting the trust is implicit, and parties should trust the process blindly which is reasonable as the technology is solid. However, the presence of the malicious actors is not accounted and could manifest into trust issues.

Capacitates all the-said threats if the protocol execution is compromised

This attributes the underlying risk with the technology that if the protocol could get compromised in some way where it would no longer hold up the promises it made, then all the business threats associated with data sensitivity can become relevant.

These shortcomings can be addressed by the incorporation of the *MPC Process Auditing* as a function of the *Governance Model* which could audit the health of the incorporated MPC technological component. This basically proves that to enable data marketplaces, the right coordination between the technological and non-technological aspects is needed which can be related to this case as a right coordination between the ***Governance Model*** and the ***MPC powered Data Exchange Service.*** This can be attributed as a hypothesis to represent in the conceptual model

Enables a Safe and Secure Data Trading with the right coordination from the Governance Model

The corresponding conceptual model representing the implication of MPC technology to the threat landscape of the data marketplaces is illustrated in the Figure 28.

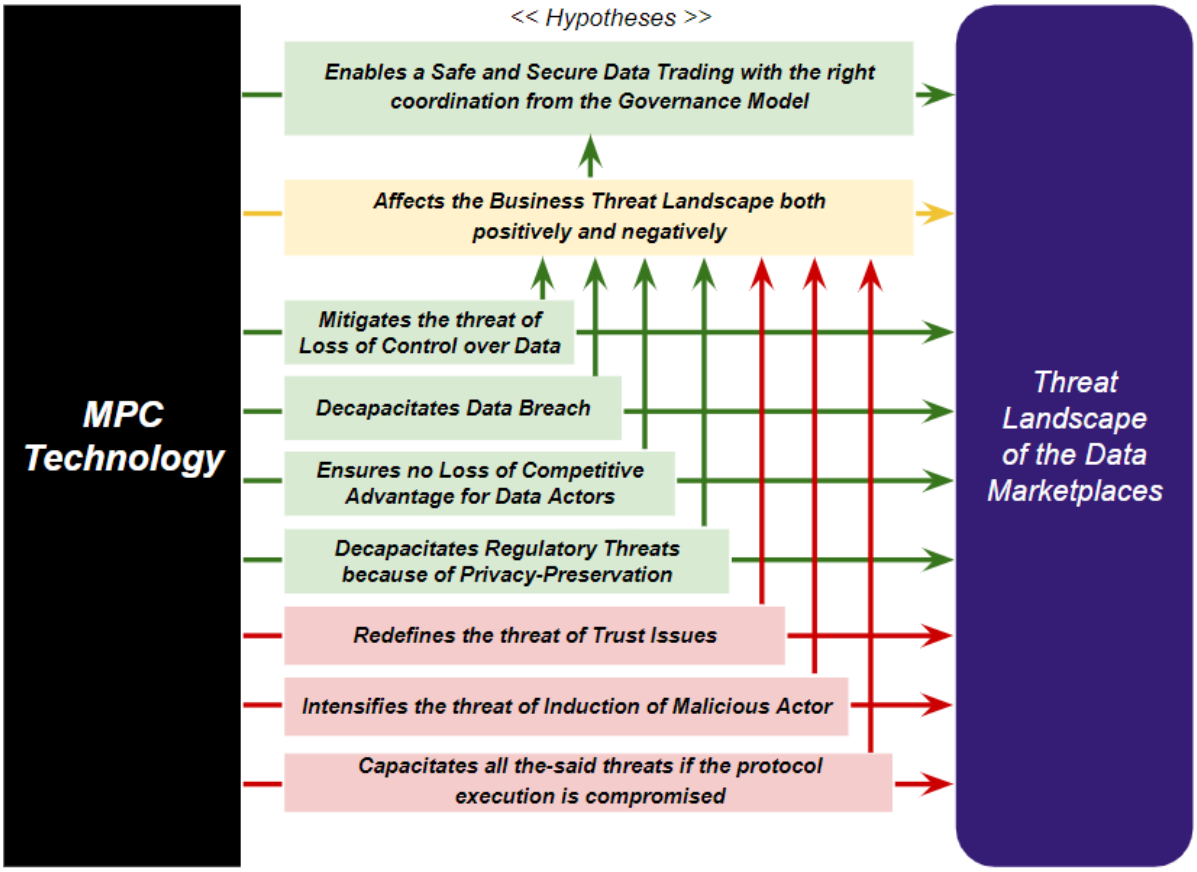


Figure 28: *Implication of MPC technology to the Threat Landscape of Data Marketplaces*

8.1.2 Implication of the Maturation of MPC technology

The previous 2 conceptual models reflect the implication of the MPC technology to the data marketplaces provided the technology has attained its maturity to be applied in the real-life applications. Hence, the 2 conceptual models can be summed up to deduce that,

"The maturation of the Multi-Party Computation (MPC) technology can enable it to be applied in the data marketplaces. This could enforce safe and secure data trading in a Security-by-Design way, thus, architecturally materialising the data marketplace platforms; and further, with the right coordination of the Governance Model, can help overcome the uncertainty around the threat landscape of the data marketplaces with respect to sensitive nature of data, commodification of data et cetera; ultimately, achieving the realisation of the business species of Data Marketplaces"

This statement which reflects the promise that the maturation of MPC technology brings to the table for the data marketplaces; signifies the theoretical framework along with the 2 conceptual models and their subsequent hypotheses. These collectively, constitute the answer to the main research question of the thesis, ***RQ.***

8.2 Sub-Research Questions

To answer the main research question, 16 sub-research questions were formulated which resulted in 2 iterations of 4 conceptual models contributing to the research objective of this thesis; The answers to these sub-research questions are presented here in the order of the 4 conceptual models comprising of both the first and second iterations for each conceptual model to explicate the change they underwent during the validation phase. The order of the conceptual models and their respective sub-research questions are as follows:

- *Pre-MPC Data Marketplace Platform (SQ1 & SQ2 (**1.0**) and SQ9 & SQ10 (**2.0**))*
- *Post-MPC Data Marketplace Platform (SQ5 & SQ6 (**1.0**) and SQ11 & SQ12 (**2.0**))*
- *Pre-MPC Threat Model (SQ3 & SQ4 (**1.0**) and SQ13 & SQ14 (**2.0**))*
- *Post-MPC Threat Model (SQ7 (**1.0**) and SQ15 & SQ16 (**2.0**))*

8.2.1 Pre-MPC Data Marketplace Platform

To obtain an architecture to reflect a generic data marketplace platform prior to incorporation of MPC technology, the following sub-research questions were formulated and were answered in the conceptualisation phase using desk research methods.

SQ1: *How to build an architecture for a generic data marketplace platform?*
and
SQ2: *How does a generic data marketplace platform look like?*

A literature study was conducted on data marketplaces with an aim to explore the phenomenon of data marketplaces involving their fundamental concepts like the definition, different features, relevant actors et cetera and also, to obtain an understanding of the architectural aspects of the data marketplace. However, the architectural knowledge was not found in the literature and hence, a framework was developed to build a high-level architecture for the data marketplace platform. This framework was referred as *HLA framework* and consisted of 3 attributes namely, Functional Requirements, Customers and Functional Components; which served as the answer to *SQ1.* Using HLA framework, from the knowledge obtained through the literature study, a high-level architecture was built for a generic data marketplace assuming it to be just a technological platform. The resulting high-level architecture reflected the answer to *SQ2.* This architecture subjected to validation through expert interviews and the following sub-research questions were formulated for the same.

SQ9: Is the *Pre-MPC Data Marketplace Platform 1.0* valid?
and
SQ10: *How do the expert insights change the architecture of the data marketplace platform from SQ2?*

The architecture was remarked to be valid for the most part, but the experts suggested further relevant improvements; which answers *SQ9*. The flagship comment was that, the data marketplace should not be viewed only as a technological platform but should be viewed as a business entity within an ecosystem. As a result, the architecture underwent significant changes to included relevant non-technological elements with respect to all the 3 attributes, functional requirements, actors in the ecosystem and function components; major addition being of a *Governance Model.* Furthermore, a new taxonomy for the data marketplace platform designs was developed in which our focal data marketplace was positioned as *many-to-many B2B decentralised serendipity model data marketplace.* This aspect was also incorporated along with a few more improvements and finally, an updated architecture was obtained which answers *SQ10.* The updated architecture is

illustrated here in Figure 29 as it reflects the updated and more valid version of the *Pre-MPC Data Marketplace.*

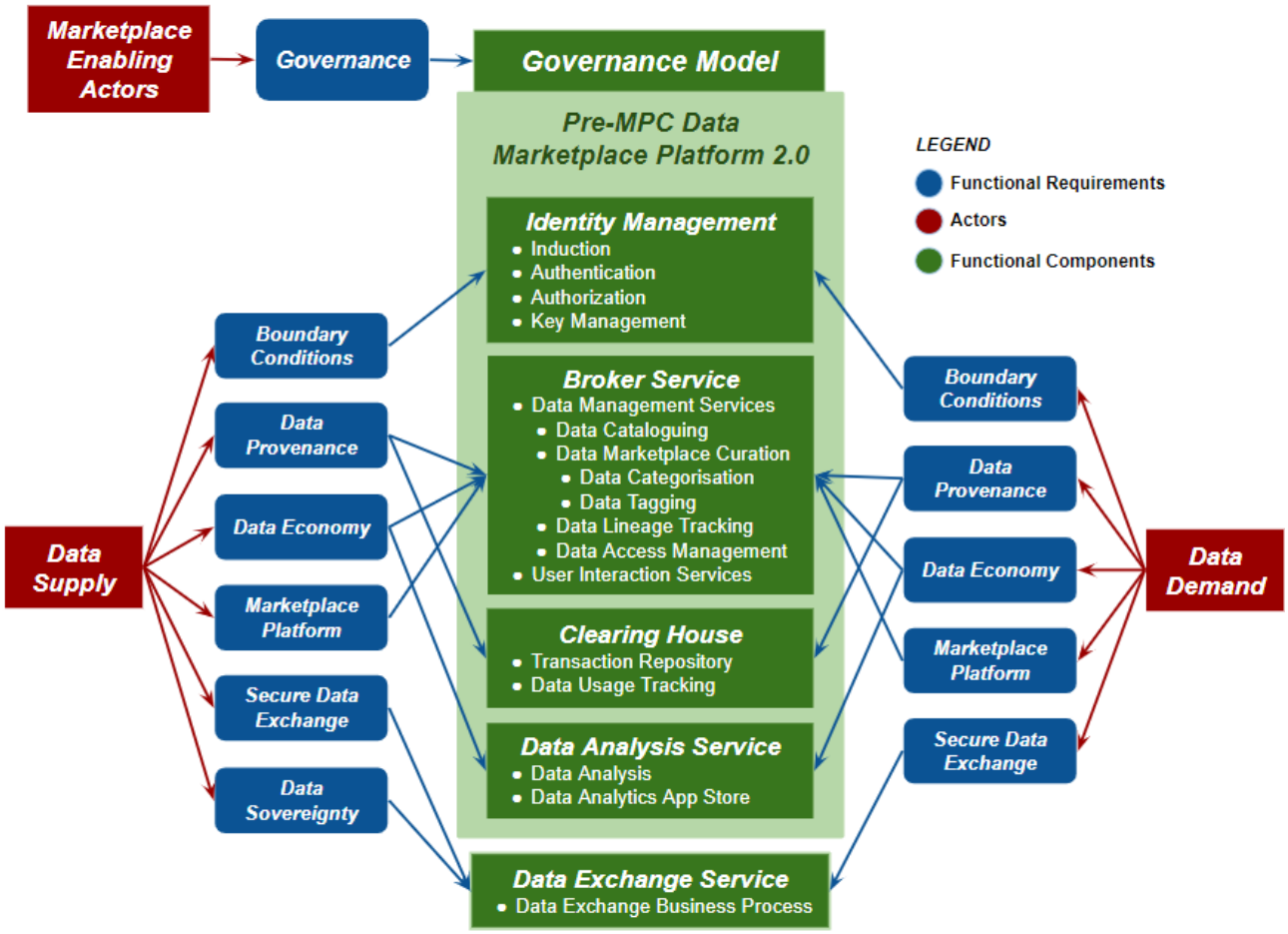


***Figure 29**: Pre-MPC Data Marketplace Platform 2.0*

8.2.2 Post-MPC Data Marketplace Platform

To understand the implication of the MPC technology on the high-level architecture of the data marketplace platform, the following sub-research questions were formulated and were answered during the conceptualisation phase using desk research methods.

***SQ5:** How to incorporate MPC technology into the architecture of the data marketplace platform from SQ2?*
and
***SQ6:** What is the effect of MPC incorporation on the rest of the architecture from SQ2?*

SafeDEED project proposal was studied to understand their conceptualisation of MPC technology and how they intend to materialise its process. It was deduced that SafeDEED materialises MPC technology with its ***SafeDEED Component*** comprising of the ***SafeDEED Primitives,*** which provides the cryptographic blocks required for building the protocol and ***SafeDEED Network,*** which provides a communication channel for the execution of the protocol. This *SafeDEED Component* provides a black box way of incorporating MPC technology for the customers who could just choose the

required function and provision and let the SafeDEED Component to build and execute the protocol. Hence, SafeDEED Component answers *SQ5.*

SafeDEED Component was integrated into the *Data Exchange Service* as they both represented a mechanism of transferring data or the knowledge inside it from the data owner to the data consumer. As a result, the platform would become decentralised where the actors can meet over the platform and the *Data Exchange Service* enabled by *SafeDEED Component* is set up ad-hoc by the marketplace outside the platform the actors to execute the protocol and share data. This was the direct effect of the MPC incorporation. There were indirect effects too which were incorporated to obtain the MPC incorporated high-level architecture reflecting the *answer for SQ6.* This idea of MPC incorporation was subjected to the validation through expert interviews and the following sub-research questions were formulated for the same.

SQ11: Is the *Post-MPC Data Marketplace Platform 1.0* valid?
and
SQ12: *What according to the experts, can be the effect of MPC technology on the architecture of the data marketplace platform from SQ10?*

The conceptualisation of the MPC incorporation through SafeDEED component to gain Post-MPC Data Marketplace Platform 1.0 was remarked as valid. However, since the pre-MPC architecture had undergone updation during the validation prior to this, the effects of the MPC incorporation were deduced for the newly obtained *Pre-MPC Data Marketplace Platform 2.0* and the changes validated and suggested by the experts were incorporated to obtain a more valid conceptualisation of the incorporation of MPC in the form of *Post-MPC Data Marketplace Platform 2.0.* The effects were as follows,

- *Data Exchange Service* is provided ad-hoc with SafeDEED component outside the platform (same conceptualisation as of *Post-MPC Data Marketplace Platform 1.0*)
- Data Analysis Service is integrated into the Data Exchange Service for MPC protocols support data analysis. As a result, the platform only contains Data Analytics Appstore.
- Since the requirement of Secure Data Exchange and Data Sovereignty are enforced technologically by MPC alone, the burden is reduced from the Governance Model with respect to these requirements.
- Eliminates key management in the identity management component as it would no longer needed in the presence of MPC.

These effects answer *SQ12* and are illustrated in Figure 30.

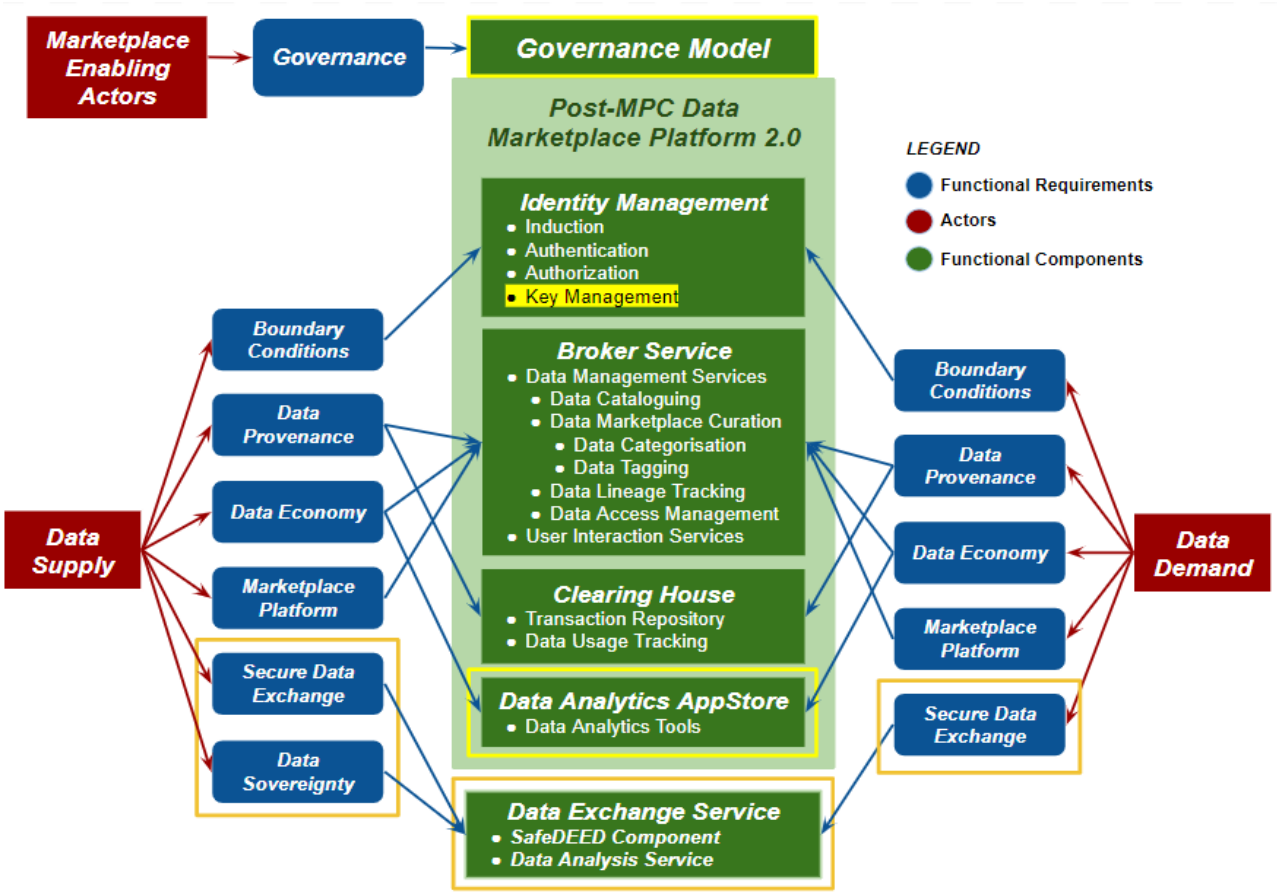


***Figure 30**: Post-MPC Data Marketplace Platform 2.0*

8.2.3 Pre-MPC Threat Model

To identify the threats associated with the data marketplaces, the following sub-research questions were formulated and were answered during the conceptualisation phase using desk research methods.

***SQ3:** How to model the threats for the architecture of the data marketplace platform from SQ2?*
and
***SQ4:** What are the threats associated with the data marketplace platform from SQ2?*

A literature study was conducted on threat modelling to understand how the process of threat modelling could be applied to the case of our high-level architecture. It was deduced that none of the threat modelling frameworks in the literature were applicable to our case of performing threat modelling at the level of business functions. Following this, HLTM framework was developed which helps in carrying out high-level threat modelling for the technological entities with high-level architectures such as ours; thus, answering *SQ3*. Furthermore, the idea of threat landscape was conceptualised which comprised of the combination of cyber threat and its business consequence to the focal entity. Using the HLTM framework and the literature analysis of cyber threats, a high-level threat model was developed comprising of the threats that affect each individual component in high-level architecture and the threats were coupled with their relevant business consequences to represent the threat landscape of our high-level architecture. The resulting threat model reflected the *Pre-MPC Threat Model 1.0* and the constituent threats served as the answer to *SQ4*.

Following this, the threat model was subjected to validation through expert interviews and the following sub-research questions were formulated.

SQ13: Is the *Pre-MPC Threat Model 1.0* valid?
and
SQ14: *What according to the experts, are the threats associated with the data marketplaces?*

The *Pre-MPC Threat Model 1.0* was remarked as invalid as the threat model did not reflect the actual threat landscape of the data marketplaces, thus answering *SQ13*. It was reasoned that the threats in the threat model were baseline threats as they were built upon the baseline architectural specification of individual components of the high-level architecture. Since these were basic threats which could be mitigated by incorporating appropriated mitigation techniques, the threat model was criticised that it did not reflect the actual threat landscape of the data marketplaces as the threats that actually hinder data marketplaces are much complex to address. The experts remarked that those kinds of threats affect the business logic at a level much higher than our low component level analysis. These threats were named as *business threats*. Further, the scope of our analysis was heightened from the low component-level to the high business level and accordingly, the conceptualisation of the threat landscape was updated to consider threats only to business logic. Further, a new threat model comprising of the business threats to the business logic i.e. data sensitivity, data handling et cetera was built. The business threats that hinder data marketplaces were deduced from the expert insights and are as follows,

- *Loss of Control over Data for the Data Owner* making the data owner reluctant to participate in the data marketplace platform.
- *Trust Issues for the Data Actors* as the trust is implied intangibly but nor established with explicit measures of mechanisms which makes hard for the data actors to trust each other in a business setting.
- *Data Leakage* where the data may not be used by the concerned party as stipulated in the contract and could be leaked to other parties, or data breach because of encryption failure.
    - *Data Leakage by Back Correlation:* special case where the anonymised data can be coupled with auxiliary data to obtain the personal information which was anonymised earlier. In addition to PII, even confidential information can be obtained by combining the data with appropriate auxiliary data.
- *Loss of Competitive Advantage for the Data Actors* where the actors owning data feel reluctant to share their data as they fear if that data might give away proprietary information which could result in the loss of competitive advantage.
- *Regulatory Threats* become relevant if the private information is involved and transacted without complying to the GDPR which may cause regulatory fines and legal complications for all the actors involved.
- Data Sensitivity: relates to the fact that all the threats are because of the sensitive nature of the data and the challenges with respect to its commodification. As long as there is physical data involved in the transaction, all the above threats prevail.
- *Induction of Malicious Actors* which is a different threat other than data sensitivity where a malicious could be inducted and he can use the platform services for his benefit while deceiving other actors with invalid participation like providing invalid data, learning about metadata of other actors et cetera.

These threats constitute the business threats which reflect the threat landscape of the data marketplaces as reflected by the experts, thus answering *SQ14.* The threat model reflects more valid Pre-MPC Threat Model 2.0.

8.2.4 Post-MPC Threat Model

To understand the effect of MPC on the threats modelled in *Pre-MPC Threat Model 1.0,* the following sub-research question was formulated and answered during desk research.

SQ7: *What is the effect of MPC incorporation on the threats associated with the data marketplace platform from SQ4?*

It was deduced that MPC technology overcomes the threats associated with data handling like data breach, privacy breach etc. by bringing about structural changes in the data exchange mechanism in a Security-by-Design way, which answers *SQ7.* While the threats of rest of the components remain unaffected resulting in Post-MPC Threat Model 2.0. This new threat model was put up for validation in the form of the following sub-research questions,

SQ15: Is the *Post-MPC Threat Model 1.0* valid?
and
SQ16: *What according to the experts, can be the effect of MPC incorporation on the threats associated with the data marketplaces?*

However, in the light of *Pre-MPC Threat Model 1.0* being invalidated by the experts, the answer to SQ7 was also discarded and thereby, even the *Post-MPC Threat Model 1.0* was deduced to be invalid; thus, answering *SQ15.* Hence, the effect of MPC technology on the newly developed *Pre-MPC Threat 2.0* was analysed by coupling it with the knowledge of Post-MPC Data Marketplace Platform 2.0. Following this, it was deduced that the incorporation of MPC would overcome the following threats.

- *Loss of Control over Data*
- *Data Leakage*
- *Data Leakage by Back Correlation*
- *Loss of Competitive Advantage for Data Actors*
- *Regulatory Threats*
- *Data Sensitivity*

These threats are mitigated as MPC enables data sharing to happen in a confidentiality-preserving and privacy-preserving way such that the physical data is never transferred to different parties but are accessed in the form of an MPC protocol which only delivers computational result to the dedicated receiver. As a result, the threats associated with the sensitivity of data become irrelevant as the data resides at the owner's site safely. However, the incorporation of MPC can introduce threats which could affect data marketplaces. These threats that exist even after MPC incorporation were formalised as,

- *Trust Issues*: The threat of data actors not participating in the data marketplace as the data actors find it difficult to trust a technological setting of just MPC protocol to handle their valuable commercial data
- *Induction of Malicious Data Actor*: The threat of inducting a malicious actor into the data marketplace as a legitimate costumer. Since the tangible trust is orchestrated by technology, the malicious actor could take advantage of this feature to gain crucial information by exploiting other actors. MPC intensifies the actions of malicious actors.
- *Faulty Execution of MPC Protocol*: This is a fundamental threat where all the promise of the MPC relies on it being functioning as expected. However, if the execution goes faulty or if the protocol is compromised somehow, then all the threats mitigated by MPC would become relevant.

This reflected the final conceptual model *Post-MPC Threat Model 2.0* which answers the final sub-research question, *SQ16*.

8.3 Contributions of the Research

The research is relevant both for the scientific community and the industry. The contributions are presented here in the following subsections.

8.3.1 Theoretical Contributions

Our research contributes significantly to the knowledge gaps in the form of the refined and validated artefacts obtained at the end of the *validation* phase. The contribution is made towards the literature of 3 focal subject areas: data marketplaces, threat modelling and MPC technology.

Firstly, the contributions made to the knowledge gaps existing in the literature of the *data marketplaces* are listed as follows,

- A new taxonomy of data marketplace platform designs was created which provides an updated classification comprising of the different platform designs containing both concept platforms and realised ones. The taxonomy refines the basic classification of Koutroumpis et al. (2017) and updates it with a variety of probable data marketplaces**.** This provides a new foundation to position different data marketplaces either during design or analysis.

- A new list of functional requirements was developed which furthers the conversation of the functional requirements from just being technological to also include non-technological aspects, helping in the comprehensive understanding of what is expected of the business species of data marketplaces to have.

- As a significant contribution to the gap in the literature involving the architectural aspects of the data marketplaces, this research presents the *High-Level Architecture* of a generic data marketplace platform**.** This architecture can act as a reference architecture for the researchers to build more sophisticated and detailed architectures for the data marketplace platforms. Additionally, the *HLA Framework 2.0* can be used by researchers to build high-level architectures for the technological entities.

- Finally, as part of our research objective, a new *Business Threat Model and a Cyber Threat model* were developed for a generic data marketplace platform which represent respectively, the threat landscape at the business level and at the low-component level. This threat models marks the first of their kind threat models for data marketplaces which contains very relevant threats promising for the research circle to explore each of them and test the individual implication of the threats to the data marketplaces; thus, contributing to the literature involving the threat landscape of the data marketplaces.

Secondly, the thesis also contributes to the literature of *threat modelling***.** The literature on threat modelling is predominantly populated with the software centric engineering methodologies to identify and model threats of entities. There has been no business focussed threat modelling. The following contributions serve the purpose of filling this void and help further research on the lines of our objective.

- A new taxonomy for threat models was created to expand the scope of threat modelling from just the low-level cyber threats to also include high-level business threats. This taxonomy goes beyond just focussing on the cyberspace and includes the analyses of threats to the business logic of the focal entity. The NGCI Apex Classification of Cyber Threat Models by Bodeau et al. (2018) is also positioned in our taxonomy.

- A new cyber threat modelling framework which operates at the business function level of information systems of technological entities was developed which goes by the name *High-Level Cyber Threat Modelling (HLCTM)* framework. This framework provides an effective threat model for detained architectures and provides a baseline threat model for high-level technological entities. Additionally, the framework provides a straight forward way to carry out low-profile threat modelling on technological entities which can be used for auxiliary tasks of researches in bigger scopes

Finally, the thesis contributes to the gaps existing in the literature of the *MPC technology* mostly associated with its business application aspects. These are discussed as listed below,

- Our research clarified the business process of the MPC technology saying that the process is dependent on the underlying use-case and hence, cannot be standardised for a platform like data marketplace except can only be provisioned in an ad-hoc sort of way. Furthermore, we explicated the application of this business process in a data marketplace platform. Thus, contributing an application for the gap involving the business application of MPC technology.

- Furthermore, we have also investigated the effect of MPC on the threats associated with data sensitivity and data marketplaces which furthers the literature explicating the advantages and shortcomings of MPC technology.

8.3.2 Practical Contributions

The resulting conceptual models of the thesis proposes Multi-Part Computation (MPC) technology as an enabler of safe and secure data trading which can be materialised with the help of *SafeDEED Component* being integrated as the process for data exchange on data marketplaces. This can help the data marketplaces to get off and achieve the true potential of fostering the data economy in Europe. This is the major contribution of this thesis in terms of its practical application. On the other hand, this formulation opens up a new business application for the MPC technology. In this way, the research contributes towards the business realisation of both the data marketplaces and the MPC technology.

Apart from the main contribution, the research also generated a High-Level Cyber Threat Model which provides a starting point for developing the security architecture for the data marketplaces with its baseline cyber threats. Furthermore, a new Business Threat Model was also developed which provides the issues that data marketplaces must address structurally to attain business realisation. In this way, both the threat models serve as necessary elements to be addressed during the design of a practical data marketplace.

A collateral outcome of the research which was discarded from our objective is the *HLCTM framework*. This framework can be used by cyber security professionals to carry out threat modelling of business entities at the level of its technological components. The framework can have 2 applications. Firstly, the framework can be used by technical professionals to carry out cyber threat modelling of an information system with defined and detailed architecture, business processes and security requirements. The resulting cyber threat model will be specifically valid to the architecture under consideration. Secondly, the framework can be used by a manager to carry

out cyber threat modelling of the high-level architectures with no implementation details where the resulting cyber threat model contains baseline threats to the focal system and provide a baseline security overview to implement that focal system in a secure way. However, the validity of the cyber threat model is sketchy and totally dependent on the assumptions made about the abstract system.

8.4 Limitations of the Research

Although the research has significant intended outcomes, they are particularly relevant in the established scope and suffer from the limitations which hinder the applicability of the results in a broader context. However, the boundaries established by the limitations for the research helps effectively in explicating the results in our scope and further guides the future research of improving the external validity of the results.

Firstly, the whole research is built on the assumption that SafeDEED achieves the maturation of the MPC technology to the extent of it being compatible for real-world applications. It is known that MPC is not technologically matured and suffers from many limitations itself. MPC is not yet scalable to be used for the real-world scale of data operations. Furthermore, since MPC protocol design is dependent on the use cases and their functions, it is still not sure if all the functions, computations and the data analysis services be implemented into MPC protocols. SafeDEED aims to overcome these very same shortcomings of MPC so that it can be applied for real-world applications. Our research developed one such real-world application for MPC in the form of its application in the *many-to-many B2B decentralised serendipity model data marketplaces*. The hypotheses developed here are valid provided the above-mentioned assumption holds good. Furthermore, on the MPC technology, the analysis is carried out based on their business relevance alone. The technical specifics of the MPC technology was not explored during the literature study as we are not experts in that area. Hence, it was a conscious decision to pursue just the business feasibility of MPC technology in data marketplaces and then the technical feasibility was validated through expert interviews.

Moving on to the methodology of the research, firstly, the literature study was carried out based on the search methodology described in their respective chapters. Though, the search was carried out to include all the relevant literature to achieve comprehensiveness, there can be a chance of overlooking or missing relevant literature due to bounded rationality and selection bias. However, provided the scarcity of the literature in the subject areas dealt here, we believe we have covered the relevant literature for the most part.

Furthermore, the purpose of validation directed us to carry out judgement sampling of experts in the subject areas. Owing to the time constraint and unavailability of experts in the holiday season, we ended up with 4 interviews with which we were able to validate our concepts at the least twice. However, we feel if we had gotten a chance to pursue a bigger sample containing more researchers and industry experts, it could have resulted in even more valid results and helping us reach theoretical saturation which was not reached.

The next limitation is with respect to the interview protocol. For the purpose of the validation of our concepts, a specification document was created comprising of all the concepts which were expected of the experts to be familiar of before the interview. Given the busy schedule of the experts, most of the experts did not get a chance to thoroughly go over the document. One of the experts even mentioned that the document was overwhelming to get familiarised as it was a lengthy affair with 14 pages even though it was purely informative purposes. This is the cost faced because of validation methodology. The validation aspect of the interviews also affected the conduction of the interviews. Even though the semi-structured nature was meant to foster discussion about the further concepts, the experts who did not possess any further expertise on

certain concepts just validated saying that the concept makes sense rather than adding anything else to the concept. Essentially, the responses of the experts were purely based on the first impressions of the concepts rather than introspected opinions. Which we would have appreciated more Furthermore, the issue along with that of not studying the specification document, made some portions of the interview less interactive and more like the expert reading the document and suggesting that the concept makes sense and moving on. If the expert had a chance to go over the document and its concepts prior to the interview, we feel it could have provided with more nuanced results.

Coming the part of qualitative data analysis, we followed the techniques of Middle-Ground Approach for the analysis. We carried out the traditional steps of Data Reduction, Data Display and Drawing Conclusions to obtain the refined concepts at the end. Because of the smaller number of interviews, we could not continue to the third stage of coding, *Selective Coding* through which we could validated the relationship between the categories and codes. Even the data reduction part was carried out by only one analyst. The data analysis should also be carried out by a different analyst to increase the reliability factor of the data analysis and thereby the results. Because of these reasons, we did not get a chance to test the *categorical* and the *interjudge reliability* (*interrater agreement*) of the analysis.

Evidently, overall limitation is that, *theoretical saturation* was not reached, and the updated list of categories and codes serves only the second iteration of theoretical concepts but not the ultimate list. Because of this reason, the internal validity of 2 resulting conceptual models of the thesis remains sketchy and cannot be judged just yet. However, this directly translates into a future work recommendation with which our research can be continued where it is left off.

8.5 Future Research

As mentioned, a lot of times so far, the research associated with the data marketplaces is relatively new. Certainly, our research presents several promising directions to pursue for future research not only in the field of the data marketplaces but also in the subject areas of threat modelling and business application of MPC technology.

First and foremost, our research can be picked up where it was left off. With more interviews and finer data analysis, the limitations of this thesis can be overcome by achieving categorical and interjudge reliability; followed by increasing internal validity. Following this, the 2 conceptual models and their subsequent hypotheses can be tested through deductive reasoning from the actual incorporation of MPC technology in real-life data marketplaces and its impact. This would complete the validation of the business application of MPC technology in the data marketplaces.

Secondly, the hypotheses generated here are for the serendipity model data marketplaces because of their generic nature. However, we have dealt with the example of MPC technology being proposed in a use-case based data marketplaces, but this is not tested either. So, this also presents a future direction to expand the application of MPC technology in all kinds of data marketplaces. Furthermore, the concepts developed in this research for the field of data marketplaces, the taxonomy of data marketplace designs and the business architecture provides further directions like surveying the actual data marketplaces to update the taxonomy, building more detailed enterprise and system architecture for the data marketplaces et cetera.

Thirdly, once the application of MPC technology in use-case based data marketplaces are tested and validated, the focus can be shifted towards developing use-cases for the usage of MPC technology. Given the data driven culture is on the rise in the Europe, inter-organizational data sharing can become a norm in many technological and non-technological industries including inter- and intra- industries. Hence, a future study can be conducted in the application of MPC

technology in every kind of inter-organizational commercial data sharing which ultimately helps fostering the data-driven economy of the Europe.

Coming to the field of threat modelling, even after our research, there still is a gap with respect to business threat modelling at the higher level of business logic of the technological entities. Basically, there are no frameworks or methodologies to carry out *business threat modelling* of technological entities. We carried this out through exploratory study by interviewing experts. However, a framework or methodology makes this task very easy and very helpful. Hence, this is a very promising future direction to pursue.

Last but certainly not the least, during our analysis of threat landscape of data marketplaces, we only considered cyber threats initially and business threats later. However, legal threats were excluded from the scope in the beginning of the thesis. However, legal threats reflect the final dimension of the threat landscape of data marketplaces; which is worth exploring for the legal community.

References

Al-Mohannadi, H., Mirza, Q., Namanya, A., Awan, I., Cullen, A., & Disso, J. (2016). "Cyber-Attack Modeling Analysis Techniques: An Overview." In *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)* (pp. 69–76). Vienna. https://doi.org/10.1109/W-FiCloud.2016.29

Arrow, K. J. (1972). "Economic Welfare and the Allocation of Resources for Invention." In *Rowley C.K. (eds) Readings in Industrial Economics* (pp. 219–236). Palgrave, London. https://doi.org/10.1007/978-1-349-15486-9_13

Bishop, M. (1991). *"An Overview of Computer Viruses in a Research Environment."* Dartmouth College, Hanover, NH, USA. Retrieved from http://www.ncstrl.org:8900/ncstrl/servlet/search?formname=detail%5C&id=oai%3Ancstrlh%3Adartmouthcs%3Ancstrl.dartmouthcs%2F%2FPCS-TR91-156

Bodeau, D., & Graubart, R. (2014). "A Framework for Describing and Analyzing Cyber Strategies and Strategic Effects", MTR 140346, PR 14-3407. The MITRE Corporation, Bedford, MA.

Bodeau, D. J., Mccollum, C. D., & Fox, D. B. (2018). *"Cyber Threat Modeling: Survey, Assessment, and Representative Framework"," PR 18-1174*. HSSEDI, The MITRE Corporation. Retrieved from https://www.mitre.org/sites/default/files/publications/pr_18-1174-ngci-cyber-threat-modeling.pdf

Brynjolfsson, E., & McAffee, A. (2012). "Big Data: The Management Revolution." *Harvard Business Review*, *90*(10), 60–68. Retrieved from http://tarjomefa.com/wp-content/uploads/2017/04/6539-English-TarjomeFa-1.pdf

Chakrabarti, A., Quix, C., Geisler, S., Khromov, A., & Jarke, M. (2018). "Goal-Oriented Modelling of Relations and Dependencies in Data Marketplaces." In *iSTAR@CAiSE 2018*. Retrieved from https://pdfs.semanticscholar.org/29ce/33d36953534defc34dcf7b01f14a7a02d0c2.pdf?_ga=2.67273746.246347384.1566747501-1794706104.1555951664

Conti, M., Dragoni, N., & Lesyk, V. (2016). "A Survey of Man In The Middle Attacks." In *IEEE Communications Surveys & Tutorials* (Vol. 18, pp. 2027–2051). https://doi.org/10.1109/COMST.2016.2548426

Corbin, J., & Strauss, A. (1990). "Grounded Theory Research: Procedures, Canons, and Evaluative Criteria." *Qualitative Sociology*, *13*(1), 3–21. https://doi.org/https://doi.org/10.1007/BF00988593

Davenport, T. H. (2006, January). "Competing on analytics." *Harvard Business Review*, *84*(1), 98–107. Retrieved from https://hbr.org/2006/01/competing-on-analytics

de Reuver, M. (2019a). MOT2312 Research Methods - 3.1. "Data Collection Operationalization." Faculty of TPM, TU Delft, Delft.

de Reuver, M. (2019b). MOT2312 Research Methods - 7.2 "Qualitative Data Analysis." Faculty of TPM, TU Delft, Delft.

Deichmann, J., Heineke, K., Reinbacher, T., & Wee, D. (2016). *"Creating a successful Internet of Things data marketplace." McKinsey & Company*. Retrieved from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/creating-a-successful-internet-of-things-data-marketplace

Dhillon, D. (2011). "Developer-driven threat modeling: Lessons learned in the trenches." In *IEEE Security and Privacy* (Vol. 9, pp. 41–47). https://doi.org/10.1109/MSP.2011.47

Eisenmann, T. R., Parker, G., & Van Alstyne, M. W. (2006). "Strategies for Two-Sided Markets." *Harvard Business Review*, *84*(10), 92–101. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2409276

Federal Financial Institutions Examination Council. (2016). *"FFIEC Information Technology Examination Handbook: Information Security."* Retrieved from https://ithandbook.ffiec.gov/ITBooklets/FFIEC_ITBooklet_InformationSecurity.pdf

Fricker, S. A., & Maksimov, Y. V. (2017). "Pricing of data products in data marketplaces." In *Ojala A., Holmström Olsson H., Werder K. (eds) Software Business. ICSOB 2017. Lecture Notes in Business Information Processing* (Vol. 304, pp. 49–66). Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-319-69191-6_4

Fu, X. (2005). *"On traffic analysis attacks and countermeasures", Doctoral Dissertation*. Texas A&M University. Retrieved from http://hdl.handle.net/1969.1/4968

Goldreich, O. (1998). *"Secure Multi-Party Computation", Manuscript, Preliminary Version*. Retrieved from https://www.researchgate.net/publication/2934115

Guszcza, J., Steier, D., Lucker, J., Gopalkrishnan, V., & Lewis, H. (2013). *"Big Data 2.0: New business strategies from big data." Deloitte Review*. Retrieved from https://www2.deloitte.com/insights/us/en/deloitte-review/issue-12/big-data-2-0.html

Hansman, S., & Hunt, R. (2005). "A taxonomy of network and computer attacks." *Computers & Security*, *24*(1), 31–43. https://doi.org/10.1016/J.COSE.2004.06.011

Hartmann, P. M., Zaki, M., Feldmann, N., & Neely, A. (2016). "Capturing value from big data – a taxonomy of data-driven business models used by start-up firms." *International Journal of Operations & Production Management*, *36*(10), 1382–1406. https://doi.org/10.1108/IJOPM-02-2014-0098

Hynes, N., Dao, D., Yan, D., Cheng, R., & Song, D. (2018). "A demonstration of sterling: a privacy-preserving data marketplace." In *Proceedings of the VLDB Endowment* (Vol. 11, pp. 2086–2089). https://doi.org/10.14778/3229863.3236266

Johnson, C. S., Badger, M. L., Waltermire, D. A., Snyder, J., & Skorupka, C. (2016). *"Guide to Cyber Threat Information Sharing", NIST Special Publication 800-150*. National Institute of Standards and Technology, U.S. Department of Commerce. https://doi.org/10.6028/NIST.SP.800-150

Jones, J. A. (2005). *"An Introduction to Factor Analysis of Information Risk (FAIR)"* (Vol. 1). Risk Management Insight. Retrieved from http://riskmanagementinsight.com/media/documents/FAIR_Introduction.pdf

Kamatchi, R., & Ambekar, K. (2016). "Analyzing Impacts of Cloud Computing Threats in Attack based Classification Models." *Indian Journal of Science and Technology*, *9*(21). https://doi.org/10.17485/ijst/2016/v9i21/95282

Koutroumpis, P., & Leiponen, A. (2013). "Understanding the value of (big) data." In *2013 IEEE International Conference on Big Data, Big Data 2013* (pp. 38–42). Silicon Valley, CA. https://doi.org/10.1109/BigData.2013.6691691

Koutroumpis, P., Leiponen, A., & Thomas, L. D. W. (2017). *"The (Unfulfilled) Potential of Data Marketplaces." ETLA Working Papers* (Vol. 2420). The Research Institute of the Finnish Economy. Retrieved from http://hdl.handle.net/10419/201268

Kraus, L., Fiebig, T., Miruchna, V., Moller, S., & Shabtai, A. (2015). "Analyzing End-users'' Knowledge and Feelings Surrounding Smartphone Security and Privacy"." In *IEEE Security & Privacy Workshops - Mobile Security Technologies (MoST)*. San Jose, CA. Retrieved from http://www.ieee-security.org/TC/SPW2015/MoST/papers/s1p2.pdf

Leiponen, A., & Thomas, L. D. W. (2016). "Big data commercialization." In *IEEE Engineering Management Review* (Vol. 44, pp. 74–90). https://doi.org/10.1109/EMR.2016.2568798

Liang, F., Yu, W., An, D., Yang, Q., Fu, X., & Zhao, W. (2018). "A Survey on Big Data Market: Pricing, Trading and Protection." In *IEEE Access* (Vol. 6, pp. 15132–15154). https://doi.org/10.1109/ACCESS.2018.2806881

Lupu, M. (2018). *"Safe-DEED: Safe Data Enabled Economic Development" Project Proposal*. KNOW-CENTER GMBH RESEARCH CENTER FOR DATA-DRIVEN BUSINESS & BIG DATA ANALYTICS, Austria.

Marback, A., Do, H., He, K., Kondamarri, S., & Xu, D. (2013). "A threat model-based approach to security testing." *Software - Practice and Experience*, *43*(2), 241–258. https://doi.org/10.1002/spe.2111

Marotta, A., Carrozza, G., Battaglia, L., Montefusco, P., & Manetti, V. (2013). "Applying the SecRAM methodology in a CLOUD-based ATM environment." In *2013 International Conference on Availability, Reliability and Security,* (pp. 807–813). Regensburg. https://doi.org/10.1109/ARES.2013.108

Meier, J. D., Mackman, A., Dunner, M., Vasireddy, S., Escamilla, R., & Murukan, A. (2003). *"Improving web application security: Threats and Countermeasures", Satyam Computer Services, Microsoft Corporation*. Retrieved from https://docs.microsoft.com/en-us/previous-versions/msp-n-p/ff649874(v%3Dpandp.10)

Muckin, M., & Fitch, S. C. (2017). "A Threat-Driven Approach to Cyber Security: Methodologies, Practices and Tools to Enable a Functionally Integrated Cyber Security Organization." *Lockheed Martin Corporation*, 1–45. Retrieved from http://ce.sharif.edu/courses/95-96/2/ce746-1/resources/root/Resources/Lockheed Martin Threat-Driven Approach whitepaper.pdf

Muscat, I. (2019). "What Are Injection Attacks." *Acunetix*. Retrieved from https://www.acunetix.com/blog/articles/injection-attacks/

Muschalle, A., Stahl, F., Löser, A., & Vossen, G. (2013). "Pricing approaches for data markets." *Castellanos M., Dayal U., Rundensteiner E.A. (Eds) Enabling Real-Time Business Intelligence. BIRTE 2012. Lecture Notes in Business Information Processing. Springer, Berlin, Heidelberg*, *154*, 129–144. https://doi.org/10.1007/978-3-642-39872-8_10

NIST. (2011). *"Managing Information Security Risk: Organization, Mission, and Information System View", NIST Special Publication 800-39* (Vol. 40). National Institute of Standards and Technology, U.S.Department of Commerce, Gaithersburg. https://doi.org/10.1108/k.2011.06740caa.012

NIST. (2012). *"Guide for Conducting Risk Assessments", NIST Special Publication 800-30 Revision 1*. National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg. https://doi.org/10.6028/NIST.SP.800-30r1

Northcutt, S. (2018). *"Security Controls."* Retrieved from https://www.sans.edu/cyber-research/security-laboratory/article/security-controls

Omotunde, H., & Ibrahim, R. (2015). "A review of threat modelling and its hybrid approaches to software security testing." *ARPN Journal of Engineering and Applied Sciences*, *10*(23), 17657–17664. Retrieved from http://www.arpnjournals.org/jeas/research_papers/rp_2015/jeas_1215_3222.pdf

Osterwalder, A., Pigneur, Y., Bernarda, G., & Smith, A. (Designer). (2014). Value proposition design : how to create products and services customers want. Retrieved from https://www.wiley.com/en-us/Value+Proposition+Design%3A+How+to+Create+Products+and+Services+Customers+Want-p-9781118968055

OWASP. (2018). "Forced browsing", Open Web Application Security Project,. Retrieved July 1, 2019, from https://www.owasp.org/index.php/Forced_browsing

Parker, D. B. (2015). "Toward a New Framework for Information Security?" In *Computer Security Handbook (eds S. Bosworth, M. E. Kabay and E. Whyne)* (pp. 3.1-3.23). Hoboken, NJ, USA. https://doi.org/10.1002/9781118851678.ch3

Quix, C., Chakrabarti, A., Kleff, S., & Pullmann, J. (2017). "Business process modelling for a Data Exchange Platform." In *Proc. Forum at the 29th International Conference on Advanced Information Systems Engineering (CAiSE), CEUR Workshop Proceedings* (Vol. 1848, pp. 153–160). Essen, Germany, 2017. Retrieved from http://ceur-ws.org/Vol-2118/iStar2018_paper_4.pdf

Ramel, D. (2016). "Microsoft Closing Azure DataMarket." *Application Development Trends MAG*. Retrieved from https://adtmag.com/articles/2016/11/18/azure-datamarket-shutdown.aspx

Roman, D., & Stefano, G. (2016). "Towards a Reference Architecture for Trusted Data Marketplaces: The Credit Scoring Perspective." In *2016 2nd International Conference on Open and Big Data (OBD)* (pp. 95–101). Vienna. https://doi.org/10.1109/OBD.2016.21

Rosenquist, M. (2009). *"Prioritizing Information Security Risk with Threat Agent Risk Assessment." IT@Intel White Paper*. Retrieved from http://media10.connectedsocialmedia.com/intel/10/5725/Intel_IT_Business_Value_Prioritizing_Info_Security_Risks_with_TARA.pdf

Sabbagh, B. Al, & Kowalski, S. (2015). "A Socio-technical Framework for Threat Modeling a Software Supply Chain." *IEEE Security and Privacy*, *13*(4), 30–39. https://doi.org/10.1109/MSP.2015.72

Schomm, F., Stahl, F., & Vossen, G. (2013). "Marketplaces for Data: An Initial Survey." *SIGMOD Rec.*, *42*(1), 15–26. https://doi.org/10.1145/2481528.2481532

Sekaran, U., & Bougie, R. (2013). *"Research Methods for Business: A Skill-Building Approach." Wiley* (Seven). Retrieved from https://www.wiley.com/en-nl/Research+Methods+For+Business:+A+Skill+Building+Approach,+7th+Edition-p-9781119266846

Simonite, T. (2016). "Technical Roadblock Might Shatter Bitcoin Dreams", MIT Technology Review. Retrieved from https://www.technologyreview.com/s/600781/technical-roadblock-might-shatter-bitcoin-dreams/

Smith, D. A. (2017). "7 Steps of a Cyber Attack and What You Can Do to Protect Your Windows Privileged Accounts", Beyond Trust. Retrieved from https://www.beyondtrust.com/blog/entry/7-steps-cyber-attack-can-protect-windows-privileged-accounts

Smith, G., Ofe, H. A., & Sandberg, J. (2016). "Digital Service Innovation from Open Data: Exploring the Value Proposition of an Open Data Marketplace." In *49th Hawaii International Conference on System Sciences (HICSS)* (pp. 1277–1286). Koloa, HI. https://doi.org/10.1109/HICSS.2016.162

Smith, J. (2018). "Data Marketplaces: The Holy Grail of our Information Age", Towards Data Science, Medium. Retrieved from https://towardsdatascience.com/data-marketplaces-the-holy-grail-of-our-information-age-403ef569fffb

Spiekermann, M., Tebernum, D., Wenzel, S., & Otto, B. (2018). "A metadata model for data goods." In *Multikonferenz Wirtschaftsinformatik (MKWI)* (pp. 326–337). Retrieved from http://mkwi2018.leuphana.de/wp-content/uploads/MKWI_147.pdf

Stahl, F., Schomm, F., Vomfell, L., & Vossen, G. (2015). *"Marketplaces for Digital Data: Quo Vadis?" Working Papers, ERCIS - European Research Center for Information Systems 24* (Vol. 24). Retrieved from http://hdl.handle.net/10419/129780

Stahl, F., Schomm, F., & Vossen, G. (2014). "Data marketplaces: An emerging species." *Frontiers in Artificial Intelligence and Applications*, *Databases*(August 2013), 145–158. https://doi.org/10.3233/978-1-61499-458-9-145

Stahl, F., Schomm, F., Vossen, G., & Vomfell, L. (2016). "A classification framework for data marketplaces." *Vietnam Journal of Computer Science*, *3*(3), 137–143. https://doi.org/10.1007/s40595-016-0064-2

Steven, J. (2010). "Threat Modeling-Perhaps It's Time." *IEEE Security and Privacy*, *8*(3), 83–86. https://doi.org/10.1109/MSP.2010.110

The MITRE Corporation. (2015). "Adversarial Tactics, Techniques, and Common Knowledge (ATT&CK)." Retrieved from https://attack.mitre.org/

Uzunov, A. V., & Fernandez, E. B. (2014). An extensible pattern-based library and taxonomy of security threats for distributed systems. *Computer Standards and Interfaces*, *36*(4), 734–747. https://doi.org/10.1016/j.csi.2013.12.008

van Bommel, P., van Gils, B., Proper, H. A., van Vliet, M., & van der Weide, T. P. (2005). "The Information Market: Its Basic Concepts and Its Challenges." In *Ngu A.H.H., Kitsuregawa M., Neuhold E.J., Chung JY., Sheng Q.Z. (eds) Web Information Systems Engineering – WISE 2005. WISE 2005. Lecture Notes in Computer Science* (pp. 577–583). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11581062_50

Watson, C., & Zaw, T. (2018). *"OWASP Automated Threat Handbook: Web Applications", Open Web Application Security Project*. Retrieved from https://www.owasp.org/images/3/33/Automated-threat-handbook.pdf

Wells, D. (2017). *"The Rise of the Data Marketplace: Data as a Service."* Retrieved from https://www.datawatch.com/wp-content/uploads/2017/03/The-Rise-of-the-Data-Marketplace.pdf

Wynn, J. (2014). "Threat Assessment and Remediation Analysis (TARA)." *The MITRE Corporation*, 14–2359. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/1016629.pdf

Xiong, W., & Lagerström, R. (2019). "Threat modeling – A systematic literature review." *Computers and Security*, *84*, 53–69. https://doi.org/10.1016/j.cose.2019.03.010

Yao, A. C. (1982). "Protocols for Secure Computations." In *23rd Annual Symposium on Foundations of Computer Science (SFCS '82). IEEE Computer Society* (pp. 160–164). Washington, DC, USA. https://doi.org/10.1109/SFCS.1982.88

Zhang, L., Yu, S., Wu, D., & Watters, P. (2011). "A Survey on Latest Botnet Attack and Defense." In *2011 IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications* (pp. 53–60). Changsha: IEEE. https://doi.org/10.1109/TrustCom.2011.11

Zlomislic, V., Fertalj, K., & Sruk, V. (2014). "Denial of service attacks: An overview." In *2014 9th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1–6). Barcelona: IEEE. https://doi.org/10.1109/CISTI.2014.6876979

Zyskind, G., Nathan, O., & Pentland, A. (2015). "Enigma: Decentralized Computation Platform with Guaranteed Privacy." *Cryptography and Security, Cornell University*, *abs/1506.0*, 1–14. https://doi.org/978-1-61208-153-3

Appendices

A Expert Interviews

A.1 E1: Reggie Cushing

A.1.1 Interview Setting

Date: 24 July 2019

Interviewer: Jeevan Kumar N

Interviewee: Reggie Cushing

Subject Area: ***SA1***: Data Marketplaces & ***SA2***: Threat Modelling

Interviewee Profile: Post Doc researcher at University of Amsterdam. Working on a project called DL4LD (Data Logistics for Logistics Data) dealing with the conceptualising of a data marketplace in the airline industry.

Relevance for the Study: Expertise in data exchange mechanisms and data marketplaces.

Interview Recorded: Yes

Comments: Reggie was not invited directly for the study but was happy to participate. We invited Willem Koeman, and he referred Reggie to us as he is working closely in designing a consortium data marketplace.

A.1.2 Interview Invitation

Dear Willem Koeman,

Greetings. Hope you are doing well.

Data Marketplaces are a new kind of data-driven businesses which have the potential to boost the data-driven economy. But they are not fully realised yet. We are conducting a study to realise their security aspects which involves cyber threat modelling of conceptual data marketplaces. We are on the lookout for experts like you to involve in our research.

I am Jeevan Kumar, carrying out Master Thesis at Delft University of Technology. The thesis is part of EU project, Safe Data Enabled Economic Development (***SafeDEED***); headed at TU Delft by professors Mark de Reuver and Tobias Fiebig. Under their guidance, I have developed a concept architecture for data marketplaces; designed a framework to carry out high-level threat modelling for the concept data marketplace; applied the framework on the concept architecture and developed a threat model based on literature. The resultant threat model consists of general threats which are intended to inform the design of the security architecture of the data marketplaces. Ultimately, I am investigating how Multi-Party Computation (MPC) protocols can be incorporated into the architecture of the data marketplaces platform and their effect on the identified threats.

I wish to conduct interviews with experts like you to understand the empirical phenomenon. As part of this, given your expertise in B2B data exchange, I would appreciate your opinions ***to validate our conceptual architecture of the data marketplace and the threat model comprising of cyber threats associated with the data marketplaces***. Together, we can gain a common understanding of the challenges that exist for the data marketplace platform to achieve full realisation.

To achieve this, I would like to schedule a ***one-hour long interview*** with you whenever you are available in the forthcoming weeks. The interview will be recorded upon your consent, transcribed and analysed to validate my work. Please grant me this opportunity as your insights will be a great addition to my work and to the ongoing research on data marketplaces. Please refer the attached document for further information. Feel free to contact me for further queries.

A.1.3 Interview Transcript

Can you introduce yourself? And tell me, what do you do? What area do you work in?

I'm Reggie. I'm working as a postdoc at the University of Amsterdam. I work on two projects, one of them is DL4LD (Data Logistics for Logistics Data). DL4LD is a Dutch project about marketplaces. It is also trying to conceptualize a data marketplace. And the other project I'm working is on distributed systems for XSK. They are not very related, but there could be some underlying similarities in the future. These are the projects I am working on.

If I understand correctly, you have expertise related to data exchange mechanisms. What are you working on specifically related to data exchange mechanisms?

We are trying to conceptualize an architecture for the data marketplace. We have a use case in mind from airline industries. And we're trying to figure out an architecture that allows for the use case to happen.

What do you think are the challenges involved in commoditizing data?

It is a big challenge, mainly, because of the technological limitations involved in trading of data. Because it is very easy to copy data and it can be done instantaneously. So, the challenges are: if we give data to someone, that someone can just quickly copy the data and take it for himself. That is where all these architectures become complicated; how to secure, how to overcome the limitations of technology and when that is not reached, how to minimize the risk of data leak. So the challenge is not a simple one. The person to whom you sell the data can just start reselling your data, because he can just have to copy it. So, how to minimise the risk and block the parties from selling your data? These concerns are related to data sovereignty. How to maintain data sovereignty in a marketplace? This is the main challenge I feel here.

How does the data exchange happen between the data owners and data consumers currently? Does it happen through one-to-one contractual obligations or if data brokers are involved? How does it happen?

My assumption would be that two parties enter an agreement and do the data transaction and a non-disclosure agreement is in place.

How do these parties find each other?

From our experience in our project, the parties within an industry come together for example, like in airline industry, because they figure out that there is a win-win situation for them to share the data. So, it starts I guess, from a motivation that sharing data is profitable for all the parties. Because no competitors will share data if he feels he is gonna lose the competition. So, this phenomenon comes into action with the idea of consortium of technical partners. Like Airline partners form a consortium, Health partners can form a consortium et cetera. They come together and agree to share data among each other under strict conditions so that everyone in the consortium can benefit. This consortium leads into the creation of the architecture.

So, basically, a use case in an industry motivates the parties to form the consortium and share data. Is that right?

Yeah. That is our approach that parties come together when they have a motivation that they can make more value. If they combined their data together, they gain more value out of it.

Are you familiar with any real life data marketplaces?

Yes. Every now and then, you come across something like that. But the term, data marketplaces is a bit overused. I came across this one called, ***Data Republic****. Its website says that it is a data marketplace. But I think at the end, they all end up being like a single domain marketplace; like a cloud provider where all the data is brought and the access is managed through authentication and authorization. For us, this is not the ideal marketplace. Our approach is more from the perspective of the architecture. Actually, we are looking for something more global, like the setup of the internet. A distributed system where providers hook up together and form one network. And the exchange will be like the internet exchanges in cities like Amsterdam Internet Exchange where the internet exchange happens. Similarly, we are conceptualizing a data marketplace where data exchanges happen with different providers hooking up to to the data exchange similar to the internet architecture. Once the idea is realised, the execution becomes much more complex.*

What do you think are the functionalities that a data marketplace should support for it to be viable?

I was looking at your blocks and I agree with these blocks and also some additional stuff. To conceptualize, we always start from a governance model. The assumption is that if the consortium of parties is coming together, then there should be an authority to manage the parties and govern the activities. So, a governance model that basically gets the parties together; establishing and agreeing on the terms and conditions, basically legal stuff. And from this, you get abstract notion contracts. The contracts may be between certain or all the parties; and the contracts contain the terms of what can be shared with who, which data can be shared using which algorithm, what computing functions can be done in this algorithm. timeframes, quality of the data et cetera. The contracts define these things. This is a requirement in addition to the blocks mentioned in the business architecture.

And then there is another criterion which is, marketplace needs trust and mechanisms to maintain trust, and here the notions of authorities come into picture. So, you can have an auditing authority. And because it is an authority, it can be independent like a Certification Authority or like the authorities that verify if your architecture is up to standard, if the architecture is implemented as designed by the architect. If you are a data provider, you need some authority that verifies if you are a legitimate provider by implementing the standards that have been defined at some point. The marketplaces should also need auditing authorities who audit the marketplace as legal teams would need to have something for litigation.

And like you have mentioned: data provenance and stuff which is embedded in your blocks. But I don't see an actor who is an authority responsible for, for example, enforcing the data provenance and

management of public and private keys. We also need to consider these as the important parts of marketplace. Perhaps the architecture can contain a block called governance which comprises of the actors like authorities, auditors et cetera. Of course, the architecture here is technological and the technological components kind of enforce the governance but technology can only do so much. At some point, there is no way to enforce technologically that someone among the parties can just copy the data and run away with it. So to limit this, you need the idea of authorities in the architecture to build a trust relationship between the parties. And we have this trust in place around the data, then with the help of the governance mechanism, there's a way to litigate legally. Basically, the most complex thing about setting up a marketplace is to find the right coordination between the technology and legal aspects and how to combine these to have a complementary effect. Because just on the technology front, there is nothing new where the use security technologies like crypto encryption etc is a directive but how it is enforced with the regulations within the marketplace is important.

Now we can dive deep into the model, the business architecture of the data marketplace platform. Here, can you reflect on each of the functional requirements of the data marketplace platform that I have conceptualised? Also, can you tell me like if these are exhaustive, or are there any other requirements from a technological perspective?

Boundary conditions:

I see you are assuming the many-to-many marketplaces, the distributed idea. And it is very high level because this comes very complex depending on how to implement barriers. Part of the governance aspect that we discussed, fits within the boundary conditions. Then, there are other notions of parties and groups in the system where a user belongs to a company and so he is allowed to access the marketplace. So, boundary conditions are in itself a topic to explore on how it can be implemented and enforced. Conceptually, boundary conditions are necessary, and they can be done in many different ways which makes it quite complex. But there should be some sort of Identity Management.

Data Provenance:

The data provenance is an important aspect. The data can be combined with the other partner and data provenance helps here. Auditing can be helped by data provenance. The transactions happening on the marketplace can be part of the audit trail. The data provenance would be part of the information gathered from the audit trail. Obviously, auditing would be much bigger task. But getting the transaction information of the data and the provenance of the data would be one part of the audit trail.

One other thing is, here you say about ownership. But it's not clear. I do not see any definition of data ownership because that might play a role in the architecture. It can be defined like for example, change of ownership means switching keys. So, you quickly realize you need some key management system along with identity management. Because I assume identity management is for users. But similarly, I guess how data is treated relies strongly on some identification of data points; how do you address that? And how do you prove ownership of the data? Technologically, we can do this with keys. So, the person who owns the private key of the encryption owns the data. The idea is similar to how wallets of the blockchain work. So, you can change ownership of the data, then there is already a risk there without the governance because data provider can still have a copy of the data. So, nothing stops him from keeping a copy of the data. And that's where then you need governance. The change of ownership concept can be part of the data inventory block here. So, basically it boils down to what does it mean to own the data and how you define it.

Data Sovereignty

It is true about the statement that you can enforce sovereignty. However, it depends on how you implement it. In centralisation, then the marketplace becomes the authority and you trust the authority. Then, the authority gives you the possibility of data sovereignty. If you tell them to destroy the data, you trust the authority that the data is going to be destroyed. And in decentralized way, you could show

that you own the data on something like blockchain but, once you get out of the blockchain, the data is copied somewhere, then you lose the sovereignty. So, if there is a possibility that data and computations stay on the blockchain, then you have sovereignty. But it is not feasible for real life applications. But it is nice of investigate. So I think it is a hard claim that in distributed design, you can always maintains sovereignty. We can research that.

Secure Data Exchange:

It is like a communication channel. But here again, once the consumer gets the data, nothing stops the consumer to do whatever he wants with the data. That is challenge that we are concentrating on rather than an external entity getting access to the data. More than the external actors gaining access to the data, the internal problems are yet to be resolved where the internal actors have to work solely on the trust over the parties involved. Again, this comes back to governance issue.

Data Governance:

Data Governance is the most important requirement and it is the requirement that establishes the legitimacy of the data marketplace. Enforcing data governance is currently done with the help of the authority who oversees all the operations on the marketplace. However, how specifically it is enforced and who enforces it, whether the actor with authority or any technology, is dependent on the architecture of the marketplace.

One way is where the marketplace provider can enforce some governance. So, the consumer's data will be with the marketplace provider. Maybe Party A says I will only trust if my data is held by some infrastructure provider, for example, Amazon.com. Let's say, hypothetically, that Amazon will provide some marketplace infrastructure where we can do some computation. And this data provider and the data consumer agree that the processing of the data is only going to happen in this controlled environment because both trusts this environment. In this case, the marketplace provider can implement this governance enforcement technologically. In this context, we talk about the ideas of what we call application architects. In this model, I don't see the concept of what we call the algorithm or computation. The way we model it is that you're sharing the data for it to be computed on or you also have the notion of what's the process of the computation. Then we have an algorithm which is basically a container for the data or whatever and the data. Then, we shall have architectures of how the algorithm and the data will come together. So, do we move the data to the algorithm or the algorithm to the data? Or do you get a third party infrastructure provider that can get the data from the two parties and the algorithm from the other party and do the computation. Then, you get the results which are the combinations that were agreed upon in our idea (use case) prior to the governance model through the contracts. And basically, it would instruct your secure data exchange i.e where to move the data depending on the contract.

All these aspects are guided and derived from the use-case. The applications, the consortium formation and the processes are decided by the underlying use case. However, the architectures are developed to have some level of programmability so that it can adapt to different use cases as there can be a potential for more use cases. But in a use case like ours which is a consortium of airline industry, the parties are fixed and there is a system with an authority who manages all these parties and the data exchanges together.

Next we can go over the functional components. You have already reflected on the Identity management that it needs a component of key management. How about the broker service? Does the way we have conceptualised makes sense to you? Is this practical possible?

Well, yeah. At some point, you need a data catalogue to describe the data to share. It could also be a distributed system. You can implement distributed catalogue libraries. Conceptually, you need the data, you need the description of the tradable objects, i.e the description of the data object. So, you have the catalogue of data objects and metadata of each data object that describes that data and then you have data at this physical location. And then that's another thing.

Can these services be automated on an online website kind of way? By services I mean the data management in the background and the user interface of a marketplace?

I don't see any reason why not. The data provider logs into a system, goes through the identity and boundary conditions and then can describe what data he can provide. Conceptually, it is possible. But implementation wise, it has its own challenges.

The next one is clearing house which is a transaction repository that contains and updates all the transaction that is happening over the marketplace. We conceptualise this with a basic idea of a database management system. What are your thoughts on this component?

There are risks involved with the management of transactions as we need it potentially for auditing. So, essentially it should be tamperproof. So, you have to think of some tamper proof database. That's where everyone has started talking about ledgers now like blockchain which preserves a trail that can be tamper proof. Perhaps not 100%, but that is better than anything to record the transactions. And you want to verify who is looking into the transactions, some entity is recording it, how to verify those entities if they are malicious or not? It becomes complex again.

Coming to the Data Inventory component, we conceptualise two designs: centralised and decentralised. Can you reflect on these designs and conceptualisation?

If this is conceptualised based on where the physical data resides, then how do you plan to materialise the decentralised design?

The motivation for the decentralised design is through the incorporation of the technologies like blockchain and multi-party computation.

We have already discussed about the blockchain. With multi party computation, the architectural implications will be huge.

Is it possible to implement the decentralised design without the use of multi-party computation?

With the multi-party computation, the system itself provides security. If you are not using multi-party computation, then I guess we can have a container with an algorithm which needs to decrypt the data. However, the container is going to copy your data somewhere else. So, the basic security is broken there. And, you have to do much more work to enforce which involves governance. With multi-party computation, you have full control and so maybe you do not need security governance. But to my knowledge, the application of multi-party computation is limited in real world applications. It would be nice to have something like that. Then the architecture would also be very different.

Without multi-party computation, can we say the decentralized design would not be truly decentralized?

Yes and No. So without multi-party computation, then you need the key management. Data is on the provider site, then you want to do compute somewhere, you need to get credential access to the data. And then how would you manage this. So, it goes to key management and need to manage credentials of the data. And this whole thing would not be needed for multi-party computation. It makes a huge difference. Then it goes back to trust. Because the data providers are going to trust you with his credentials, and some would not trust and would not give the keys.

Basically, without MPC, there is a change of ownership. And hence, there is no complete guarantee of security, but in the case of MPC, there is no change of ownership and that is more secure way of doing it. Is that right?

Yes, yes, that's how I see it.

Can you reflect on the last two components: Data Exchange Service and Data Analysis service?

The exchange service looks okay, and we discussed it earlier, I guess. With data analysis service, if it is hosted by the marketplace provider, then there is necessity to model some credential management to identify marketplace provider. If it is hosted by a third parties like cloud providers, then there has to be a mechanism to validate and verify the legitimacy of the external entity to ensure that for example, the cloud providers are trustworthy.

This brings end to the business architecture of the data marketplace platform. We can now move on to the threats associated with the data marketplaces. Can you reflect on the validity and usefulness of the high-level threat model?

Yes, when I went through this, the only issue I found with the threat model is that the threat model already assumes certain implicit architectures. And the model could quickly change if you just take a different architectural design. But also, I see the threat models as kind of services themselves. You have higher level threat models like of data leakage. For example, you're sharing data under some contract which restricts its usage and so on. The threat models where the data can be correlated with another data set which can lead to some leakage. For example, you try to anonymize the data, for example, by removing some items like personally identifiable information. One of the threat models is that this anonymized data can be correlated back to the identities if it is combined with other appropriate datasets. So, these are the high level threat model cases that still need to be addressed in the setting of a data marketplace. And what you have are low level security threat models of your components which are fine. But the higher ones are the more complicated ones like, what can technology enforce? And what can technology not enforce? So, the set of threats here in your threat model, technology wise, you can provide hundred percent security. But another set of threats that are difficult to identify, for example, this kind of higher-level data leakage. That cannot be easily prevented.

Other than this, what I see is that there's this implicit architecture in the back of the threats. For example, in identity management, you're already assuming identity database. So, it took this architectural decision. But obviously, this architecture can be designed in some other way. Then it goes in a different track. Maybe I am using x509 certificates as identities which wouldn't be a database. In that case how would your threat model change? So, maybe you can define and describe the detailed architecture, then the threat model has a foundation for validity. Because, now there is a missing layer between the concept and the threat model. I see the missing architecture. And when you say database and the SQL, these are concrete technologies. So, it depends on the architecture. Again, I see a website and password attacks, here you need a concrete architecture for password-based authentication or certificate-based authentication. So, basically these underlying architectural assumptions behind the threats need to be described before. So, describe the architecture in one way.

On a high level, the threats that we need to be concerned with respect to data marketplaces is the data breach. Is that right?

Not only data breach or leakage. When you talk about the data sovereignty, you have threats that can break data sovereignty. How you lose ownership of the data or if you agree how the data is going to be used? The threats that you have considered here which are low level. And then you have threats which are high-level like data leaks. For example, if you provide a database of all the people in the Netherlands, and you try to anonymize it, and you say you can use this data, then there exists a threat model that someone would get this anonymized database and correlate it to the identities. The threat model referred runs as an application which corelates the anonymized data to the identities.

And also, the architectural decisions you make and the applications you use in these functional components can expose the data to the threats. Some of the threats can be high-level as we discussed

earlier, and these threats need special mechanism to be mitigated. The problem with the threat model here is that even though it is a good approach and it is serving your motivation to develop baseline threats and baseline security requirements for the data marketplace, these threats are low level and the chosen mitigation techniques address these threats but not the higher-level ones like data leakage in a complete way. Because here the mistake is that you can say all the possible threats here can be mitigated with the mitigation techniques and you have 100% secure system. By that, the data should ideally never leak. But that is not true practically. That could be a risk of the system. I doubt if 100% prevention is ever possible. Usually, you will deduce that you are minimizing risk to certain extent.

So, you can think of a layering system. The threats that you have described are of the lower layer with attacks on websites. And the threats can become complex as you go up. They become more difficult to mitigate. So, then you cannot mitigate but only minimise.

How do you think these high-level threats can be modelled? Is there a framework or these can be figured out merely out of knowledge and experience?

The threats that I mentioned, I came across them mostly from talking about data sensitivity. When I talk about data sensitivity, they come into layers. And everyone is classifying data sensitivity into different categories like top secret, confidential etc and then talk about what are the implications of an attack on a different sensitive data? So, for example, aggregate data can be taking an average over the data. So let's say you have a database of all the people in the country. That's very sensitive data. But if you take an aggregate, like, the average height of the people, the output data, that's an aggregate, is of much lower sensitivity. All these concepts are discussed in data science. Lot of material around just the leakage. And with a lot of AI being done now, there's all this back correlation of census data. So, maybe you could talk about the risk assessment of the application itself.

The problem with sensitive data is that it can relate logically. So architecturally, you have everything secure. But the algorithm that's being applied on the data can itself be a threat as it can cause data leak. i.e if the data is not properly anonymised, then the leaked data itself can be sensitive even without back correlation. So, in this example, even how to anonymize can be a big issue.

Another example in a context is having health records, that's a big thing, because you can get for example, the MRI images. You can say the MRI images itself can be processed and with tracking preferences, determine that the MRI image itself is already identifying people because it becomes like a fingerprint. Although it's anonymized and it doesn't mean anything to get the MRI images, but by correlating it with image processing and other dataset, the sensitive data can find the home it belongs. Something like these are the risks associated with the data marketplaces.

Do you think MPC can be kind of a solution to these issues?

I think there is no hard solution for this. What you can do is how to minimize the risk. So, the mitigation techniques like anonymization, you can minimize this. I don't think there's a way to mitigate these risks to hundred percent. That's where you come back to governance issue because eventually, when there is some data leak, you have the auditors and everything that you can litigate legally.

A.2 E2: Mihai Lupu

A.2.1 Interview Setting

Date: 29 July 2019

Interviewer: Jeevan Kumar

Interviewee: Mihai Lupu

Subject Area: ***SA1***: Data Marketplaces

Interviewee Profile: Research Coordinator of ***SafeDEED***. Working closely with research partners to develop the enabling technologies for B2B data sharing like MPC, Data Valuation etc. Also working closely with Data Market Austria in its conceptualisation.

Relevance for the Study: Experience in materialising a real-life data marketplace, Data Market Austria.

Interview Recorded: Yes

A.2.2 Interview Invitation

Dear Mihai Lupu,

Greetings. Hope you are doing well.

Data Marketplaces are a new kind of data-driven businesses which have the potential to boost the data-driven economy. But they are not fully realised yet. We are conducting a study to realise their security aspects which involves cyber threat modelling of conceptual data marketplaces. We are on the lookout for experts like you to involve in our research.

I am Jeevan Kumar, carrying out Master Thesis at Delft University of Technology. The thesis is part of EU project, Safe Data Enabled Economic Development (***SafeDEED***); headed at TU Delft by professors Mark de Reuver and Tobias Fiebig. Under their guidance, I have developed a concept architecture for data marketplaces; designed a framework to carry out high-level threat modelling for the concept data marketplace; applied the framework on the concept architecture and developed a threat model based on literature. The resultant threat model consists of general threats which are intended to inform the design of the security architecture of the data marketplaces. Ultimately, I am investigating how Multi-Party Computation (MPC) protocols can be incorporated into the architecture of the data marketplace platform and their effect on the identified threats.

I wish to conduct interviews with experts like you to understand the empirical phenomenon. As part of this, I would appreciate your expertise and opinions ***to validate our conceptualisation of the data marketplace platform, the threats associated with it and validate our ideas of using MPC methods in the data marketplace platform***. Furthermore, ***we need your ideas also if there are additions to our work***. Together, we can gain a common understanding of the threat landscape of the data marketplaces and the application of MPC to prevent them.

To achieve this, I would like to schedule a ***one-hour long interview*** with you whenever you are available in the forthcoming weeks. The interview will be recorded upon your consent, transcribed and analysed to validate my work. Please grant me this opportunity as your insights will be a great addition to my work and to the ongoing research on data marketplaces and MPC methods. Please refer the attached document for further information. Feel free to contact me for further queries.

A.2.3 Interview Transcript

Can you tell a little bit of yourself and what do you do in SafeDEED? What are the things that you handle?

*I'm the Studio Director for research in Data Science and in **SafeDEED**, I am the scientific coordinator. I was the one who put the proposal together. I see the project is very important because it brings together the technology from the cryptology side, from the more general data analytics side with outlet into the data evaluation. That's the technology part together with the business part which is what Mark has been studying and working on, and also with the legal part which is what our colleagues from KU Leuven working on. My role is to keep an overview of the project. I have a colleague here who is working on the data deanonymization and anonymization part which is one of the tasks in work package 5 which deals with privacy; where other colleagues from the KNOW centre at TU Graz are developing the multi-party computation technology.*

I came across that you are working closely with Data Market Austria which is a data marketplace developed as a consortium with partners coming together out of a use-case to share data. Our concept is a many-to-many data marketplace where data providers can show their data and find appropriate partners to create use cases on how they can share data among each other. Are these two designs entirely different? How do these designs relate to real world data marketplaces?

I think when you develop such a thing, you have this chicken and egg problem where they develop technology, but you don't have the use cases or if you have the use cases, you do not know the appropriate technology to use. What we do in Data Market Austria is to go with it from all sides. We have 7 work packages that sort of do what you've described generally as the market, what a general data market is. And then we have 2 work packages 8 and nine, that look at use cases and sort of went in parallel to develop the use cases and now sort of, are bringing those use cases to the data market technology. There's an overlap in people between the various work packages so that they haven't been working completely independently. I think this is the challenge to that we face in general in data markets. Bringing the data on such a market will not happen organically unless there are use cases and use cases are difficult to foresee unless there is already an existing data market.

The data marketplace that I have conceptualized is a little different as it works in an eBay kind of way where parties who don't know each other can just upload their proposition and hope for finding the right partners. What do you think of this idea?

That is the same idea of the data market. So technically, this can happen. The problem with a data market like eBay is that people will just not release the data even for costs or not. We're not talking about free data or open data. Because of a variety of reasons. The data culture is not yet there. A data market like eBay seems not to be functioning right now. We wouldn't first try it right. There are several other things to get right currently.

Now we can move to the actual part of the interview which is to validate the conceptualisations. If you can look at the business architecture of the data marketplace platform, you can find the functional requirements and their respective functional components. Can you reflect on this?

The boundary conditions as described here helps and is required. However, it doesn't guarantee the intended requirement from the description here. That is good phrasing.

In data provenance, the phrase "change of ownership" is difficult to place. I have never heard of a situation in the data market where the data changes its owners. Typically, what we see in Data Market Austria and the use cases we have is that you have a data owner and a data provider. And that entity

maintains what you call as data sovereignty because we people don't call it the ownership or property over the data. What exists here are licenses. And the licenses may include redistribution or whatever. But in terms of data sovereignty, it is maintained by one entity, and there is no possibility of changing this, at least not in the Data Market Austria. In Data Market Austria, there is no process involved where the change of ownership happens. You can perhaps interview experts working on data ownership to make this concept clear.

Coming to data sovereignty, as a requirement, this is fine. But the statement in which you say with technology the provider can protect the legality of the data, is a very weird thing to say here. The data is either legal or illegal. So, I don't understand how you can protect the legality of the data. Perhaps this is where MPC will play a role to say not to be worried about the data being misused by the data consumer. Currently, blockchain will not prohibit you in this case. Blockchain is a Distributed Ledger. It does not protect you from the misuse of data. In the end, it might tell you who misused the data, but it doesn't protect you from misusing the data.

Secure data exchange seems fine. But by data subjects, if you mean data providers and data consumers, then you should call them data actors. Because by data subjects, I would understand it as entities about which the data is.

Data Governance seems to be the combination of secure data exchange, data sovereignty and data provenance. Because it says, it is a way of having mechanism for maintenance and management of data. So, keeping track of it essentially, i.e. traceability, which is data provenance, and data usage you can relate it somewhere to data sovereignty, where you keep it in blockchain to know who is using your data.

Data Economy by the description seems fine. Data Exchange Platform seems to be everything we just discussed till now. I am not sure what this particular requirement does on its own. If it is intended to just complementary requirement from all the other requirements, then that's fine.

In terms of requirements, they seem reasonable for a data market.

Do you think the requirements are exhaustive? Or are there any other requirements here that are missing?

Well, I think they are exhaustive actually in the sense that they are generic enough that you can call them exhaustive. It has covered all the bases in a general sense like, you want your data to be secure; the exchange to be secure; how the data should be used according to your rules i.e. data sovereignty; trace where it comes from; trace who has access to it. So, in that sense, they are exhaustive. But they are general.

Now, we can move onto functional components. Can you go over these functional components and reflect on them?

The identity management looks fine.

In case of broker service, one thing we do in Data Market Austria is that we separate the meta data and the data. The data itself can reside somewhere else and it doesn't matter but the metadata is centralised. Our broker services completely rely on the metadata. Here, you address it as (meta)data which is confusing.

The idea here is to compare the designs of the data marketplace before and after the incorporation of the MPC. So, in case of the design before MPC, we are considering both the designs: centralised and decentralised. In centralised version, the central entity manages both data and metadata where as in decentralised version, it manages only metadata. To express this in a generalised way, we have used the term (meta)data. It is a design issue actually.

Well, it's a big design issue. But that is fine as long as these issues are explained in the design. But the market can be decentralized without considering MPC. Data Market Austria is decentralized is not bound to MPC. And the decentralized version does not give more guarantees than a centralized one in terms of data misuse. As long as somebody has access to the data, they can write a function on it, that copies the data, and then, subsequently misuse it. That is their intention.

Moving on, no problem in the frontend features.

There was another thing that I wanted to point on data tracking. You say data tracking covers both data lineage and data usage. And I think those are very different. Data lineage is essentially what has happened to the data, like different versions, what has been added or how the owner or the sovereign of the data has changed it etc, which is kind of things you want to trace. And the data usage is about who has had access to the data, and who has obtained access to the data and whether they have downloaded or not. I think those are two different components. That should be distinct. They are distinct in our architecture, and therefore, I think they should be distinct as well. In that sense, the data usage aspect is more part of the clearing house, as you call it, because the clearing house essentially keeps the track of transactions like when somebody has access to the data. So, clearing house seems to me the place to put the data usage tracking.

Moving on to the data inventory, I'm confused here. Obviously, again, I'm trying to map everything to what I knew. So, you have a broker service that has a data management backend feature. You have an inventory, which seems to be the same thing here. So, I'm not sure how do you see the difference between the data inventory and the data management of the broker service.

The difference according to me is that data inventory is just a database. It houses the data that is being uploaded by the data owners or the data providers. On the other hand, broker services are like the processing component which takes care of all the processes that are happening on the data marketplace platform.

Okay, then the broker service makes sense. But then you say it has data management. And that seems to be done by the data inventory. So, perhaps you can rename the data inventory as data store if you just need to signify the data there, or you can remove that component as its features comes under the data management feature of the broker services.

I'm now trying to wrap my head around the data exchange service. I am thinking about how this happens. You have the broker and you find the data you want, let's say you somehow agree with the clearing house that you have access to it by having bought it or whatever. And then you have access to it. So, all of this has been done by the time we reach the data exchange service. Then the data exchange service is essentially a download link which you can make it secure over SSH. It seems like a very basic thing to have. Like saying that the internet is part of the data market which is true. But it's like an underlying condition. Without it, we wouldn't even be talking about anything. If I understand correctly, the data exchange service is actually seen as the network, the connection between the two endpoints.

The data analysis services part is also what we see as very important, the data itself is useless unless you can analyse it. Here, in our case (DMA), we have a mix of data analysis services that are centralized, so they are running on the data market platform itself. And then there are these services that are provided by third parties that essentially you can think of as an app store having a set of programs that can be executed on a data set. Then I'm okay with this part.

Moving on to the high-level threat model, what do you think are the important assets associated with data marketplace platform?

Well, identity is always important. You don't want to allow somebody to impersonate somebody else. So, obviously it has to be identity and then the data itself. This is what most people are afraid of when we talk to them about the data markets, is that they will lose control over their data and that the data

will be out in the wild, even if it is behind the paywall. Somebody else will pay for it and then they will release it. They give the example of obviously, movies or whatever. They are all behind the paywall, and then somehow, they all ended up on some BitTorrent site. And therefore, having that in mind, they see that as a big blocker to release data from behind a paywall.

When I came across the challenges with respect to setting up a data marketplace and why the data actors are reluctant in adopting this kind of model are basically the loss of ownership of the data and the threat of data breach. So, these are the two common threats that I could find. And the same are reflected in the business consequences in the model. Apart from these, what do you think are the threats experienced by data marketplaces?

Regarding the loss of ownership, again, I've never heard of it in this format. People don't phrase it like this. People phrase it in concrete threats they see to their business. So, in production and manufacturing, producing data from the machines has the potential danger of a competitor reverse engineering their processes. For instance, it can be like, they have a special process that they produce some plastic at a certain temperature, which makes it better or more stable. And then if they release sensor data from the machines about energy consumption and operation times, then based on the energy consumption, perhaps the competitor will be able to determine the temperature they're using in the process. This is an example in the industry and manufacturing. In all other sectors like Banking, Telecom or the Health, of course, the problem is with the regulations. They are afraid at some point that the data will be deanonymized and therefore, they will be facing fines for having released personally identifiable information.

Coming to your tables here, this is an area where I'm not very much an expert. Here, it goes into very much details about the kind of attacks that that people might exert on different components of the data markets. I have given you the information of what I think are the threats on a general level to the data marketplaces.

Can you now reflect on our conceptualisation of how MPC can be applied in our data marketplace platform and the effects it brings about?

This has been the great promise of MPC. The one that is shown in figure which is the interactive approach. This is the great promise of MPC. You will only be releasing data on which only a specific function can be applied. Which is you would never be able to see the data itself, but you can get results out of the data. So, that is the great promise.

The processes that we have mentioned here of MPC are different business processes which enable the data transfer between the data actors. So, with these processes, the data can be transacted from the data provider to the data consumer in a secure way without having to reveal the data itself. On these lines, in our design, MPC comes into picture only with respect to the data exchange service. MPC basically modifies and improves the processes involved in the Data Exchange service in a way that the data can be transacted in the most secure way. In addition to this, MPC eliminates the need for data store on the platform as the data now resides on the site of the data owner. So, data inventory also gets eliminated and it just becomes metadata inventory. Apart from these, no other components undergo change because of MPC. What do you think of this proposition?

You should interview our colleagues at KNOW Centre who are developing this technology. Because they have different models of how the constellation might happen.

I don't see why it needs to remove the need for data store on the central node on the central platform. Technically, you never have the need to store data on the central platform. You can always end up with distributed system because ultimately, the data transfer happens between two endpoints on the internet. Wherever the two endpoints are, it doesn't matter. All the management of the endpoints can happen on the central node. So, whether one endpoint provides data encrypted as in MPC or not

encrypted as in traditional way, I don't see why this changes anything. Perhaps data exchange service can be the component and data transfer from point A to point B is the process involved in this feature. In that way, even if the data is transferred to the platform, it can be stored there in a distributed system, which makes the use of terms confusing. So, the problem is with the phrasing. You can say the data transfer is a traditional form of data exchange and this can be replaced with data exchange service through MPC which is safe. Apart from that, I think you're right. In this design, the data exchange service is the most effective one and I think the rest do not get affected. The rest of the components can have same processes.

The next idea is that after the incorporation of MPC, then the online data marketplace platform becomes a platform for the data actors to find each other create appropriate use cases. And then, the MPC process can be established between them in an ad-hoc sort of way with the help of SafeDEED Primitives and SafeDEED Network. That is the end result of this conceptualization. What do you think of this proposition?

That's fine. Again, I will go back to what I said at the beginning. I think it is a bit naive to think that people find each other and then find the use case. I think what happens is some innovator has an idea that he wants to do so and so and looks for the data. However, somebody can just browse through the data and after browsing the data, comes up with an idea. Maybe that happens as well. I'm not excluding that. But I see it less likely. I think people have the idea of what they want to do and then they look for the data. That's more how I see the more successful use cases.

I think the data marketplace has value propositions that are not relying on this sort of serendipity kind of business development which are related to the ease of access to data. So, let's say that I do have an idea. But I do know that whichever company has the data. However, if I were to go outside of the data market to get the data, it would be complicated. I would have to go and find somebody there and contact the legal department and it's all sort of a big process. Then to integrate them, you would have to agree on a standard on how to communicate, how the data is et cetera. But with the data market, it facilitates all of that significantly and that is the promise of the data market concept. It does facilitate serendipity because you are allowed now to use the brokerage service to browse through the data and sort of just let your mind wander about the uses of potential different data sets. But in a more concrete way and in a more realistic way, the application or the benefit of the data market is the fact that if you have an idea already, it just makes it easier for you to implement that idea.

So, you can say that when an innovator comes up with an idea or a use case, data marketplace is a good place for him to search for the data that he needs. Is that right?

Yeah, and then once he finds the data, the data marketplace also simplifies the process to just use the data. Theoretically, what we will be doing is when the market would be a functional, what you would foresee is that, instead of having to make phone calls and send contracts on post and sign the contracts and whatnot, he would just click his way through and get the data within half an hour perhaps. Basically, the data marketplace orchestrates the process of data transaction and exchange in the most efficient way. It is all about the efficiency of the data exchange. It is not that it makes things possible that were previously impossible. Previously, it was possible, but it just took forever.

A.3 E3: Swati Manocha

A.3.1 Interview Setting

Date: 30 July 2019

Interviewer: Jeevan Kumar

Interviewee: Swati Manocha

Subject Area: ***SA2***: Threat Modelling

Interviewee Profile: Manager in the domain of Cybersecurity and Privacy at EY. Provides auditing and security assessment services to business clients.

Relevance for the Study: Expertise in threat assessment and security frameworks

Interview Recorded: Yes

A.3.2 Interview Invitation

Dear Swati,

Greetings. Hope you are doing well.

Data Marketplaces are a new kind of data-driven businesses which have the potential to boost the data-driven economy. But they are not fully realised yet. We are conducting a study to realise their security aspects which involves cyber threat modelling of the concept data marketplaces. We are on the lookout for experts like you to involve in our research.

I am Jeevan Kumar, carrying out Master Thesis at Delft University of Technology. The thesis is part of EU project, Safe Data Enabled Economic Development (***SafeDEED***); headed at TU Delft by professors Mark de Reuver and Tobias Fiebig. Under their guidance, I have developed a conceptual architecture for data marketplaces; designed a framework to carry out high-level threat modelling for the concept data marketplace; applied the framework on the concept architecture and developed a threat model based on literature. The resultant threat model consists of general threats which are intended to inform the design of the security architecture of the data marketplaces.

I wish to conduct interviews of experts like you to validate the work mentioned above. Your expertise and opinions are crucial for me to understand the empirical phenomenon of cyber threat modelling and to infuse those insights into my work. Together, we can gain a common understanding of how efficiently cyber threat modelling can inform the security decisions of unrealised information systems like data marketplaces.

To achieve this, I would like to conduct a ***one-hour long interview*** with you whenever you are available in the forthcoming weeks. The interview will be recorded upon your consent, transcribed and analysed to validate my work. Please grant me this opportunity as your insights will be a great addition to my work and to the ongoing research on data marketplaces and cyber threat modelling. Please refer the attached document for further information. Feel free to contact me for further queries.

Thank you in advance.

A.3.3 Interview Transcript

Can you introduce yourself and tell a bit about your experience and what you do?

I am Swati Manocha. I am a Manager in the domain of Cyber Security and Privacy at EY. We provide auditing and security assessment services to our clients. I am involved in projects where we assist our clients to implement the cyber security frameworks which includes controls and processes. We operate at the governance level or business level and offer solutions.

Can you explain the process of security assessment? Who is involved in the process? And in what capacity? Like, is it a management personnel or the technical person.

It is a combination of both. When we help our clients implement the frameworks, we do need an involvement from the management as well, because as I mentioned it is a framework at the business level, it has kernels into different levels. So, management plus the middle management as well come into the picture and then, wherever needed, the technical people. So, when you think of cyber security, you can think of a lot of mitigating controls. If you take an example of access management, in those cases, we would have technical people involved who will actually implement these things in an organization.

Here, we are trying to gauge the security aspects of the data marketplaces with no implementation details basically with a high-level overview of business functions. Do you think this is possible? Is it possible to deduce the required security policies or technologies just by analysing the business functions?

Well, it's partially possible I would say, because there are different aspects when it comes to cyber security or information security. And you have to think of it in 3 principles, which are confidentiality, integrity and availability, that also entails privacy sometimes. So, if you look at just the business functions, you can try to identify what kind of processes or procedures they have defined, what kind of requirements they have laid down towards the cyber security and privacy. So, that's partial. And of course, about how people are aware towards these threats, and how are they trained, but then the other aspect is how these requirements are enforced and implemented into different systems. So, it could be that when you're looking at the business organization, you can look at, let's say, their access management policy which is defined that could be strong, but if it is not implemented, per say, then it's not helping with preventing threats.

Before carrying out the security assessment of a focal entity, what kind of information do you expect from the clients to provide you to carry out the assessment?

The scoping. Usually, it will be about what is the scope and the type of security controls that they have in place. Security assessment for me is a very broad term, it could be looked into any aspects of the work that we do. We try to understand the objective of the organization, and how mature they are? Of course, the IT controls and the business controls that they have in place. And based on that maturity, we try to understand how vulnerable they are towards the different threats. Our work depends basically on the maturity of the security controls in place.

We have defined this prerequisite information as the context of threat modelling. Some of the attributes that we have identified from the literature are: scope, which includes organizational level, our business function or information system. Then purpose, which involves risk assessment, system design, security technology profiling. And then approach which involves asset-centric,

system-centric and threat-centric. Do you think there are any other attributes like these to establish the context of security assessment?

I would see also trying to understand the interested parties for the organization, and if they have any requirements towards the security within the organization. So, it could be regulators, let's say for a data marketplace. They may have a requirement that they are meeting certain security level within this kind of industry, or it could be the customers that may have a specific security requirement, the company. So, I guess all these things should also be taken into consideration at the beginning of the security assessment; to be able to identify the right scope of the assessment. I think these all factors would also encompass the context that you say.

Our assumption is that the high-level threat models that we are focussing on in this project, help in judging the high-level security overview of the abstract system that we intend to develop. So, the high-level threat model lends itself and guides the development process of the focal system. What are your thoughts on this? Do you think these kind of high-level threat models useful?

Definitely, it is useful. I think it is a good start to have a set of threats that could be applicable to the type of organization or any focal entity. I am looking at your high-level threat model and here it seems like you are focussing only on IT threats. I think you can include other threats like regulatory threats – noncompliance to the regulations, or natural disasters; because they are just as important on an organizational level. You can also include the threat of not properly securing and protecting personal information also cyber threat.

Now, we can get into the validation of our work which includes: the threat modelling framework and the threat model. Can you go over the Business Function centric Cyber Threat modelling framework and reflect on it if it is applicable or useful?

This looks good to me. Usually, when we do a risk assessment, we look at the assets, threats and we also look at the vulnerabilities. I don't know if that is something that you have explored. Vulnerabilities basically means as I mentioned earlier, like what is the current level of organization when it comes to the security aspects or security controls. So, when you consider the threat of the man in the middle attack, then you also look within the organization on how vulnerable this organization is to a man in the middle attack. And that gives you the consequence. May be the one that you mean by business consequence here, I guess.

The gist of the framework is that, we are breaking down the abstract system into components and their respective business functions. To identify the IT assets, we are assuming certain architectural concepts here, for example, if you consider like a website can be an IT asset. And we are identifying threats with respect to that asset. And we are trying to identify the consequence on the business because of the threat on this asset. This can be a contradiction as we assuming the architectural aspects before the implementation of the system. What do you think of this approach of assuming some system's architectural concepts before even implementing it?

Since you are just doing the generic research, I think that will give you some generic direction towards the security in this domain. But of course, following this, maybe your research would need to get customized according to the organization's requirements. For example, I think you should consider the actual architectural concepts in place for some data marketplace and then do threat modelling. The threats you would find in that kind of approach will be more valid as it is based on the actual architectural components in place.

What do you think about the computer security properties: confidentiality, integrity and availability? Are there any other properties that are applicable?

Since the nature of your research is generic, CIA holds good here, I think. Although, there are new extensions of these properties like accountability, privacy. However, all these can be encompassed within the umbrella of CIA.

We have applied this framework on the business architecture of the data marketplace platform and obtaining the threat model that you can see. Can you reflect on the framework itself? How relevant is it and how valid is it compared to the frameworks you use in the industry?

Yeah, this looks relevant to what I have seen in the market. And as we have discussed earlier, I usually see vulnerabilities in the frameworks. Other than that, I think the framework is strong I can say.

Can you reflect on the high-level threat model now? Can you comment on each of the cyber threats that we have included in this model?

Threats: Identity management

The threats look relevant to me. With respect to identity management, I think you include access management. I guess you have termed it as authorisation here. Password management is also a prominent part and I see you have included password attacks which addresses that I guess.

Threats: Broker Service

The threats seem reasonable to me. Perhaps, you can also address the threats that affect the availability of the server. Something like Denial of service attacks to the server as you have included server as the supporting asset here. I'm looking at the supporting assets like applications carrying out data management services. What if one of the services or servers is not available, the implications like the impact on the continuity of the services. The threats here can be any defect in the hardware leading to the server being down. Like a threat of system failure. You can include threats related to these issues. Or a malicious insider trying to break the server. Coming to the user interface, the threats here look fine. I cannot think of any other threats.

The threats in the rest of the sections of the threat model look fine to me. I feel the threat model if good and comprehensive.

What do you think of the quality of the threat model compared to what you have come across in the industry?

The threat model is quite relevant to what I've seen in the market. So from that perspective, it makes sense. And the framework also looks strong. And I guess you have done your research on the type of threats and mitigation techniques and everything. So all in all, it looks good. Probably, you can represent the threat model in a more interpretive way so that it is easier to look at it and discuss about it. The threat model here seems too intense in the first look. But other than that, from the perspective of research, this study looks good to me.

A.4 E4: Sebastian Ramacher

A.4.1 Interview Setting

Date: 16 July 2019

Interviewer: Jeevan Kumar

Interviewee: Sebastian Ramacher

Subject Area: ***SA3***: MPC Technology

Interviewee Profile: Researcher in ***SafeDEED***. Works on the implementation of Multi-Party Computation (MPC).

Relevance for the Study: Expertise in MPC technology and its applications.

Interview Recorded: Yes

A.4.2 Interview Invitation

Dear Sebastian Ramacher,

Greetings. Hope you are doing well.

Data Marketplaces are a new kind of data-driven businesses which have the potential to boost the data-driven economy. But they are not fully realised yet. We are conducting a study to realise their security aspects which involves cyber threat modelling of conceptual data marketplaces. We are on the lookout for experts like you to involve in our research.

I am Jeevan Kumar, carrying out Master Thesis at Delft University of Technology. The thesis is part of EU project, Safe Data Enabled Economic Development (***SafeDEED***); headed at TU Delft by professors Mark de Reuver and Tobias Fiebig. Under their guidance, I have developed a concept architecture for data marketplaces; designed a framework to carry out high-level threat modelling for the concept data marketplace; applied the framework on the concept architecture and developed a threat model based on literature. The resultant threat model consists of general threats which are intended to inform the design of the security architecture of the data marketplaces. Ultimately, I am investigating how Multi-Party Computation (MPC) protocols can be incorporated into the architecture of the data marketplace platform and their effect on the identified threats.

I wish to conduct interviews with experts like you to understand the empirical phenomenon. As part of this, given your expertise in MPC, I would appreciate your opinions ***to validate our idea of using MPC methods in the data marketplace platform***. Furthermore, ***we need your ideas also on how MPC methods can be integrated into the data marketplace platform***. Together, we can gain a common understanding of the value proposition of MPC methods in the realisation of data marketplace platforms.

To achieve this, I would like to schedule a ***one-hour long interview*** with you whenever you are available in the forthcoming weeks. The interview will be recorded upon your consent, transcribed and analysed to validate my work. Please grant me this opportunity as your insights will be a great addition to my work and to the ongoing research on data marketplaces and MPC methods. Please refer the attached document for further information. Feel free to contact me for further queries.

Thank you in advance.

A.4.3 Interview Transcript

What is Multi-Party Computation? In the umbrella of privacy preserving technologies, how is it different from other technologies?

The basic idea of Multi-Party Computation is as the name suggests, to bring different parties together to compute something on their inputs. So, one can always think of this can be done, if the parties all sent their data to a trusted authority. The trusted authority computes whatever function you want to compute, you get the result back and, and in this case, only the trusted authority learns all the data. The other parties don't know anything except, they get the result. And they know their own data, but they don't know the inputs of the other parties. This is only true to some extent. If you have to respond, you have your input and the function is very simple, then you can deduce something from the output. But this has to be considered when one develops kind of a concrete functionality. But in general, the idea here is that you only learn the result and nothing else. And the goal of MPC is to essentially get rid of this trusted third party that you don't want to have. Or that you usually don't have it at all. And what MPC allows you to do is any function or functionality that you could compute with this trusted third party in mind, can be transformed into a protocol that is computed only by the involved parties without a trusted third party. And you can still achieve the same security guarantees, which means that, again, you only know your own data, you know the result, but you don't know anything else about the inputs of the other parties.

This has some interesting applications, especially if you want to do computations on data, which are kind of sensitive, in the sense of, has some private data in it. One of the nice examples that have been implemented is that it is not legally possible to combine data from health insurance companies with hospitals. They can't just share their databases and then check how often is a person sick? Or are there any other trends like people with higher education get sick less often. It could all be statistics that if you could combine the data, you could check that right. But you can't, because the hospitals have very strict rules about what they can do with the data; the health agencies or even the government have very strict rules on what they can do with the data. And they are not allowed to just send the database to the other party or to some trusted third party, which then could combine it and compute something on it. But now, what would MPC allows you to do is that you can still compute this statistic on the data. Because the data never leaves your premises in a way that the other party can decrypt it. They only get random data that looks random where they can't use anything from it but in the end, you could still run this computation on it and you get this result.

This is what was done by a company called Chairman. they also have a system specifically for these statistical analysis of databases in this MPC setting. You have a database on patient data, you have some other database on patient data and you can run statistics over the databases for example, every hospital in the country, right, you can check if people are sick in certain areas of your country and to what extent are they sick, if it depends on a global trend. So, you can combine the appropriate databases and it gives you useful information; for your research context but also maybe for governance context, and so on, so that you can focus on your health politics in a certain area in a different way to reduce certain sicknesses or something like that.

And these are statistics and is just one of the many functionalities that you can compute with MPC. It was also used to run an auction system where the bids stay private, until the final bid is decided. You can also think of systems that compute too many different functionalities.

What is nice is the property that you have this guarantee that even if you work with the plain data that you have on a customer or a patient, or whatever personal data that the company has, that you can use the data without leaking any of the personal identifying information that is contained in the data. This is really useful, but you still have to keep in mind that the function needs also have the property that if you have the input and the function output and you don't learn anything about the other data. This still needs to be ensured in some sense. If you have that, then it can be useful. In this case, you wouldn't even need any anonymization techniques, because you don't have to send this data in plain, but it always is kind of randomized in a way I have asked, that the others can't learn anything from it. And this is kind of the appeal of MPC in this context.

Can you touch upon the business process of the Multi Party Computation? How is it materialised with the help of SafeDEED Primitives and Network component?

It is not really fixed how this will all work out. But in general, the idea is that our use-case partners find use cases where it makes sense that they interact with other companies. For example, we can have information that is sensitive in some sense, it does not need to be privacy critical data, but it can be some data internal to the company that should stay confidential because it would reveal something critical about their business in general. You don't want to leak how well your company's running, or too much details about it. So, what we are trying to find are use cases with the use-case partners where this makes sense.

One of the ideas here is to run something called a private set intersection protocol, where the main idea is you have your customer data like the list of your customers and another company also has their customer data; And you can sell products that would be interesting for customers that are subscribers in both companies, for example, you could think of targeted marketing offers. But now the question is, how do you find the customers that you share in common. You can't send the database to the other company to see what we have in common. But with private intersection protocol, which is a special type of MPC protocol, you get exactly this information.

And the idea with a marketplace can be like you can say you have this data, you're interested in using this data and running some analysis together with other companies, and the marketplace kind of is this platform where you bring those companies together. But they have to run this protocol only between them. Otherwise, everything would run through the marketplace and the marketplace would get the data and would be liable if the data is compromised because the data marketplace might get hacked or descended to the wrong company or some other scenario. So, the data marketplaces can be kind of the place where companies find each other, but then they run a protocol together among themselves. And the ***SafeDEED component****s would then run on both of the company sites and they can interact with each other. You have the* ***SafeDEED*** *code running on company A's server and on company B's server and they will then talk to each other. This would be the idea.*

Before applying the MPC protocol, you say it is necessary to know the use case consisting of what (the datasets) and how (the analysis function) datasets are combined and computed. Why is that needed? What is the concrete relevance here?

Yes, it is very critical that you know the use case beforehand, because it will influence a lot of the decisions you have to make: to choose the correct protocol, to set everything up, to run the protocol efficiently.

For example, with the private set intersection protocol, this is already a selection where you know, the information that you're looking for and so you apply the protocol. So, you have to identify that use case beforehand. Otherwise, you won't gain anything from the MPC because you can't just do it locally and run some different analysis and check what you get of it. You'll have to think beforehand what you want to achieve from it.

What do you think about the two MPC processes (interactive and non-interactive) we have conceptualised?

In the homomorphic encryption case, you would still need to think about what data you want to encrypt and send. But then you have a little bit more freedom, if you set it up correctly in the first place. So, if you want to have this data encrypted and sent to the aggregator, then you can tell the aggregator to run many number of functions on that encrypted data. And he can do that without fetching the data. If you do this in MPC, since its interactive, you will have to run the full MPC protocol again for a different function. In this aggregator process with homomorphic encryption you can prepare for this scenario

where you might want to run different kinds of analysis and functions without always contacting the data provider again and again.

Even in the case of homomorphic encryption, it is necessary to know the use case beforehand. So, the problem that the homomorphic encryption case is solving is the with respect to the availability of the data actors. Is that right?

Yes, this is one of the issues it is trying to sort. And it is a little bit different. The main difference is really that the data provider only has to provide the data once and then everything else can be done without the data provider.

Can you reflect on the two processes we have mentioned here (synchronous and asynchronous)? How accurate are these processes?

Yeah, the first one matches quite good. It matches with what we've had in one of our deliverables as one of the scenarios. Here, you can really play with all the different combinations that you can have. In this schematic, you have 3 data providers that are all active and compute something and somebody receives output. And then you have different variations. Everybody receives the output or just one receives the output, somebody that is not involved in the computation receives the output. And you can also think of nodes that only do computation without providing actual data. You just add some instances on the cloud service for example if you worry that some nodes go offline. But you don't have to have data. The idea will only be computation component.

In the second, I'm not sure if it is the best idea to call this MPC as well. It's kind of right. I mean, the data aggregation fits quite well; Homomorphic encryption fits quite well; but it lacks this interactive component that you always have in your PC. And so, maybe just call it homomorphic encryption or homomorphic aggregators, something like that. But the picture of fits quite well for the process.

Can MPC functions be carried out on encrypted data that is encrypted out of homomorphic encryption?

Yeah, if you look at the details of some of the MPC protocols, they actually use homomorphic encryption. It can be part of the protocol. There is a mix of different approaches where you have homomorphic encryption sometimes and sometimes you don't. For example, you can define an MPC protocol that compute something on homomorphic encrypted data. The question is how useful that is because you could just do that without MPC on the encrypted data anyway.

For example, when you use the homomorphic encryption approach, you have to think of which public keys to use for encryption and who has the secret key. If the aggregator and the receiver are different entities, then it is somewhat clear. You encrypt the data for the receiver; the aggregator gets the data and combines the data; and the aggregator only sends the aggregated result precise to the receiver. So, the receiver never sees the inputs, everything is fine. But if you, if the aggregator and the receiver are the same person, then it's a little bit tricky. Then you encrypt for the receiver. But now he's also the aggregator, so he sees all of the encrypted data and could decrypt the inputs. So, the system doesn't really work.

But what one could think of is that, for example, this decryption of the aggregated result is done using an MPC protocol where nobody has the full secret key, but every party has only a part of the secret key. And one party alone can not decrypt some ciphertext. They all always have to work together to decrypt it. And then you have again, this property that the receiver wouldn't learned any of the inputs.

So in the second case, if the receiver and the aggregator are the same person, then MPC can help the situation and it can enable the data sharing. Is that right?

Yeah, exactly.

To clarify how MPC can be applied in data marketplaces, MPC does not come into the picture of the data marketplace platform directly, but once the use case is generated after establishing relationships between the data actors, we can create a communication channel, which is powered by MBC protocols, between them and thus enable the data sharing. Is that right?

Yeah, exactly the data marketplace itself will never see the data, it only establishes connection between its customers. And then, they have to run the protocols on their own.

After understanding these processes, I tried to apply this knowledge of MPC in the data marketplace platform diagram that I have prepared. What I deduced is that MPC affects only the data exchange service and data inventory components and does not affect any other component. It affects data exchange service as MPC enables the data sharing process. And since the data will not be shared over the data marketplace platform, there won't be a necessary for data inventory. So, MPC eliminates data inventory component as well. Is there any other effect MPC can have on these components? I came across something related to MPC affecting data analysis. Any thoughts on that?

The data analysis can be run with the help of MPC protocols. Since the MPC component are run by the data owner, the consumer and the aggregator, data analysis services are also run at these actors' sites. The marketplace itself can just be a way to find the interested parties who you want to exchange data with. So, in this picture, the data exchange service, the data analysis services, I think would move from the platform to the sites of data owners, consumers and aggregators.

Is the data analysis also dependent on the function that needs to be computed?

Yeah, if you want to run data analysis, then you would have to define this as the function of the MPC protocol. And then you can run analysis and on top of the data.

Based on this knowledge, MPC will overcome the threats with respect to data breach that can happen on the data marketplace since the data doesn't reside on the platform anymore. We are trying to analyse everything from the perspective of data marketplace and not the data actors. In that regard, since MPC will enable the data marketplace to operate in a decentralized way, there won't be a risk of data breach happening on the data marketplace. Are there any other threats with respect to data sharing which are addressed by MPC? other than the data breach and the privacy aspects?

Yeah, I think the threats that you mentioned for data analysis service wouldn't be an issue anymore because this would be something that is run by the users of the marketplace.

But this broker service which is probably where you connect people to each other is something that is a lot more critical. Because if you connect parties with each other that don't work well together, then you have your reputation loss probably there. So if you are a malicious actor, and you use the marketplace, and you do computations with everyone; but you always just make up all the data, then it doesn't look too good from the perspective of the marketplace. But this is something that you probably also have in a classic scenario where the data analysis would be run on the data marketplace itself. If you do the analysis with invalidate or faulty data on the platform, then you would have the same issue. With MPC, this probably just moves to a different part of the system.

If the data exchange and data analysis services move to the users' site, then broker service will be the crucial component. However, MPC does not directly address the concerns or the threats of the broker service, right?

No, this stuff will not be affected. There will be an issue of finding correct parties and how you get them to interact with each other. But this is something that is out of the scope of MPC. MPC enters the scenario when you have established a use case and the relationship that these 3 companies will compute something, but not before that.

Can MPC help in the activity of how to identify the function that needs to be applied, and in turn help in finding the appropriate parties?

No, I don't think so. On the platform, you would probably say, I have this type of dataset and would want to use it. But I am interested in some statistical analysis or something like that. This is something you can't do with MPC. In finding the data or identifying what kind of data you have, and what you're interested in, you'll have to think about that as a company. You probably have to check who the other party is and if I am interested or allowed to work with them? And what kind of data do they have? I think this issue can't be solved with MPC.

Does MPC introduce new threats? What are the concerns with the incorporation of the MPC?

This is a good question. In the MPC protocols, we always assume that everybody's honest. But the parties might be curious; which means they compute everything as they should, but they might want to find out a little bit more about the data. So, if you run a protocol that is secure only in honest-but-curious setting; but one of the parties isn't honest, then you have an issue. This is one of the issues that can come up. So, you'll have to make sure that you select a protocol that is really secure for your setting that you're interested in.

And the other case is with malicious parties; where a party might exist that sends you garbage, who doesn't follow the protocol at all, tries to manipulate you into revealing data that you shouldn't reveal resulting in an information leak. But those protocols you always have the boundary conditions like how many parties need to be honest and how many parties could in theory be malicious? which means that if you now set up MPC between different customers, there could be an issue if they are not properly checked if they are honest, then you suddenly have a system where most of the other parties that you work with are malicious and then you can't run any of the protocols? And the question here is: What do you do? Do you set up contracts between customers and the data platform which could indicate if they behave maliciously? And you somehow find out that they acted maliciously, then they have to pay some fine? Or do you find a way as a data platform to make sure that there are always enough honest parties that you are still in the boundaries of the security guarantees. And this is probably one of the issues you don't have in the classical sense because in the classical one, you never directly interact with another party and it is all done by the data marketplace. So, this could be one of the new issues.