Winter 12-10-2021

The Application of Design Thinking on Evaluating a User Self-Service Data Analytics/Science Platform

Aheeka Pattnaik
apattnaik@my.harrisburgu.edu

THE APPLICATION OF DESIGN THINKING ON EVALUATING A USER SELF-SERVICE DATA ANALYTICS/ SCIENCE PLATFORM

Aheeka

Harrisburg University

GRAD 699

Acknowledgements

A massive appreciation to my Texas family and my mentor for all their support.

Thank you, Dr. Tamara “T” Peyton, for speaking to me through the lens of a Business Analyst, for guiding me to understand my thesis statement better through the lens of the UX world, and for helping me never switch from British English.

Thank you, Dr. Richard Wirth, for all your valuable inputs and direction on my thesis proposal, pulling in Dr. T when I was a little lost, and helping me put a proposal and ultimately a thesis out that I’m proud of and is representative of what I wanted to achieve in this space.

Abstract

This thesis is aimed at utilising design thinking and the first half of the double diamond framework to i) set-up a research and select the appropriate participants, ii) gather requirements and define user personas from those eligible participants, and then iii) define the framework for evaluating a user self-service data analytics/science platform. Derived from the author's own experiences, both as a Business Analyst (BA) and Citizen Data Scientist, with no-, low-, and code-based data analytics and science platforms are being implemented for enabling user self-service analytics – for users who are completely new to the space of data analysis and science *as well as* those who are experienced analysts and data scientists across a variety of industries and global regions – and there has been a need to outline an enablement process for this space. Through this research, the current state of the marketplace is researched, analysed, and evaluated alongside user research carried out on the feasibility and applicability of a UI- and UX-centric framework for ensuring human-centred design. A literature review showcases the benefits of human-centred design for humans when it comes to usability and techniques for such an application in various other fields. The key aspects of this research are to understand the users' capabilities, needs, and wants, then categorise those users into personas, analyse and segment the requirements, create functional and non-functional requirements for platform capabilities, and then, ultimately, provide an evaluation framework for any organisation and/or individual looking for a user self-service data analytics/science platform by carrying out a pilot research study on ten (10) participants.

Keywords: *Data analytics/science, data analytics/science platform, data analytics/science platform evaluation framework, design thinking, human-centred design, user centred design, user self-service analytics, Artificial Intelligence (AI), Machine Learning (ML)*

Table of Contents

List of Abbreviations

Acronym	Full Form	Acronym	Full Form
AI	Artificial Intelligence	*LNCS*	Lecture Notes in Computer Science
BA	Business Analyst	*LOB(s)*	Line(s) of Business
BI	Business Intelligence	*ML*	Machine Learning
BSA	Business Systems Analyst	*MNC*	Multi-National Corporation
BUs	Business Units	*MS*	Master of Science
CPT	Curriculum Practical Training	*PaaS*	Platform-as-a-Service
DA	Decision Analysis	*PO*	Purchase Order
DL	Deep Learning	*R&D*	Research & Development
DSML	Data Science Machine Learning	*RPA*	Robotic Process Automation
EDA	Exploratory Data Analysis	*SaaS*	Software-as-a-Service
HCD	Human Centred Design	*SDLC*	Software Development Lifecycle
HCID	Human Centred Interaction Design	*STLC*	Software Testing Lifecycle
IIBA	International Institute of Business Analysis	*SQL*	Structured Query language
IS	Information Systems	*UI*	User Interface
ISEM	Information Systems and Engineering Management	*UX*	User Experience
IT	Information Technology	*UXD*	User Experience Design

List of Figures and Tables

Diagrams

Figures

Chapter 1: Introduction

1.0 Problem Statement

The research is aimed to understand how user self-service analytics in data science can be achieved through a user experience centric focus to enable business self-service for both technical and non-technical users who are either looking to get into or expand their skills even further in the space of data analysis and science through design and UX thinking concepts.

The complete research problem statement is: The application of design thinking and UX design concepts for determining and evaluating a user self-service enabling data analytics/science platform from the lens of the user personas for its capabilities.

In line with the research problem statement, here are a couple of the deeper questions that were sought to be addressed:

i. What are the skillsets of users in the space of data analytics and science? This skillset understanding will help in categorising the needs and wants of users, and then eventually translating it into a data science platform that not only supports all those skillsets, but primarily factors in how to make itself usable for the purposes of business self-service in the space of analytics.

ii. How to ensure that the focus of a platform one will be selecting is on the adoptability and usability by users, as opposed to pure functionality? Factors such as UI and UX play a role in a platform decision analysis, which is why it is imperative that it is considered for a platform that provides the capabilities to support a variety of users.

Furthermore, as part of this pilot research was focused on understanding, eliciting, elaborating, and documenting responses derived from the following five (5) questions:

A. How do individuals currently carry out data analysis/science techniques and engage with the tools/technologies they use?
B. What are the current tools, technologies, and services and how are they used?
C. What are the challenges faced by individuals in carrying out data analysis/science?
D. What are the features and capabilities that individuals would like to have to support their data analysis/science challenges?
E. What are the standards and/or expectations for data access, performance, and volumes?

1.1 Background

The author of this paper is an Information Technology (IT)/Information Systems (IS) Business Analyst (BA), who also wears the hat of a Data Analyst turned burgeoning Citizen Data Scientist, splitting their time between requirements analysis and gathering, a.k.a. understanding the needs, wants, and requirements of stakeholders on both Business (non-technical) and IT (technical) sides of the house, as well as data analysis.

Both components go hand-in-hand, with the latter consisting of various data analysis techniques, such as analysing, extracting, cleaning, exploratory data analysis (EDA), understanding, and visualising for presentation purposes to my stakeholder, and then utilising the results of that for the former. From that combined exposure and experience, the author has learnt that there are not only data learnings about data quality issues, anomalies, and patterns/trends that can be assuaged as a result of such analysis, but, even broader than that, they have seen first-hand need for such analysis to be carried out by others as well – be they technical users or not.

With the ever-growing quantity of data, a.k.a. Big Data, the author has experienced an increase in requirements pertaining to users, specifically non-technical Business users, wanting the tools and technologies that contain functionalities to enable them to do the data analysis. According to Gartner (Columbus, 2021), as showcased in Figure 1.0, the data analytics field consists of two parts: i) Traditional Business Intelligence (BI) and ii) Advanced Analytics. While the author used to solely occupy the traditional BI space initially, they have now started living in the latter over the past three (3) years as well. Their Business stakeholders, too, have been in the same boat and are venturing into the latter space in order to make sense of their data with the objective of data-driven analysis and decision-making.

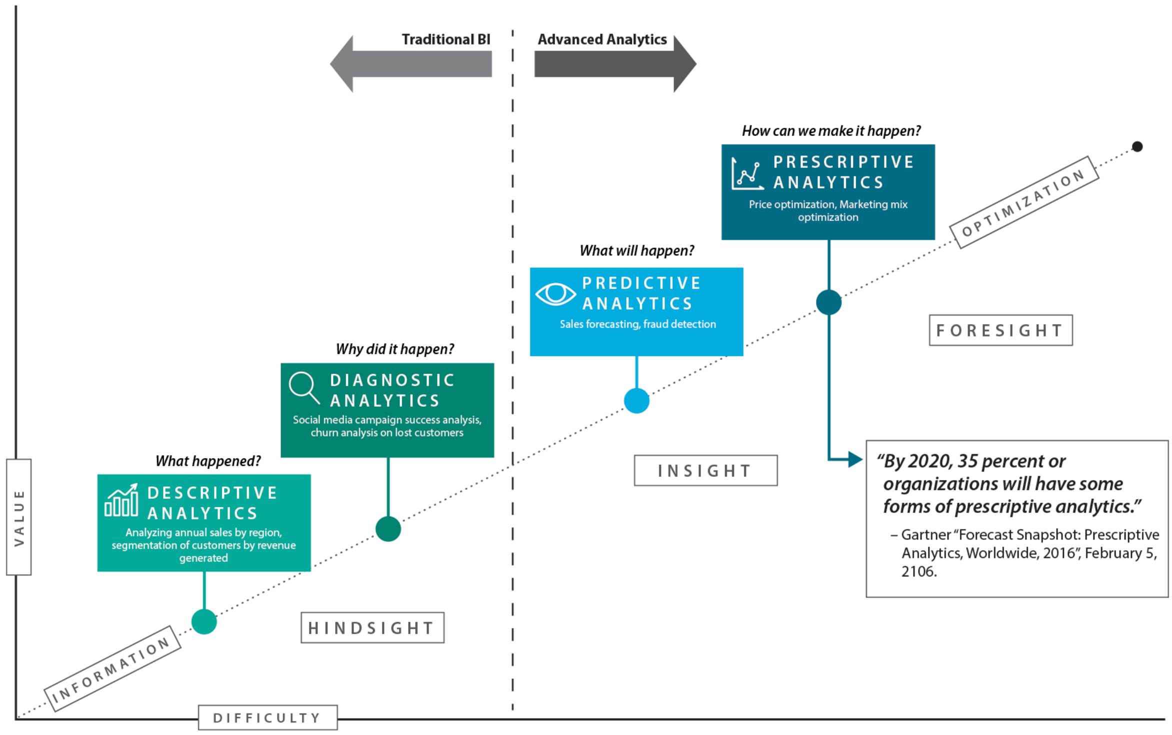

Figure 1.0. *Gartner Analytics Model Maturity (AMM)* (Columbus, 2021).

The advanced analytics space utilises data science techniques, which consists of understanding the business wants and needs (a.k.a. requirements analysis and gathering), data mining, data cleansing, data exploration, feature engineering, predictive modelling, data visualisation, and optimisation/refinement, as illustrated by Figure 2.1.

According to Gartner, "[b]y 2020, 35 percent of organi[s]ations will have some form of prescriptive analytics" (Columbus, 2021) and that is the trend that the author's client is working towards adopting. The skillset of their Business stakeholders wanting to be in this space ranges from those who are a) entirely inexperienced with data analysis/science but wanting to learn, b) experienced with data analysis through Excel, not unfamiliar with data science concepts, but wanting to break into the space, and c) experienced in carrying out data analysis and science through languages such as Python, R, and SQL. The question the author has encountered ultimately is how to support such varying levels of skills without *forcing* any skillset group to move up or down, a.k.a. being able to carry out data analysis and science through code via programming/scripting languages or not.

With the various lines of businesses (LOBs) that fall within the automotive organisation, there are a multitude of use cases that Business users want to perform data-driven analysis and decision-making on, with a few examples highlighted below:

a) Determination of part sales forecasting/predictions by season,
b) Determination of part wear and tear,
c) Customer segmentation and likeliness of vehicle purchase,
d) Pre-owned vehicle, demand forecasting and inventory management,
e) Vehicle price optimisation,
f) Warranty analysis,
g) Driving behaviour and pattern recognition,
h) Robotisation by means of robotic process automation (RPA),
i) Smart cities and smart(er) vehicles build and research & development (R&D),
j) Manufacturing fault prediction and preventative maintenance, and
k) Autonomous driving and detection through image analysis.

The primary goal of this research is to propose a framework that will allow users wanting to perform data-driven analysis and decision-making to consider a user-centric focus when evaluating and selecting a data analysis and science platform that supports a variety of user skillsets.

1.2 Relationship to CPT

As part of the author's Curriculum Practical Training (CPT), they are consulting at an automotive multi-national corporation (MNC) as the Lead IT BA, liaising between IT and the business branches to improve the quality of IT services by analysing business needs, wants, and requirements to design and/or modify business/IT systems.

Being a BA requires a Computer Science background, including skills like technical understanding and SQL querying in tandem to critical thinking and analysis capabilities, at the undergraduate level to be able to evaluate Business and IT processes for efficiency optimisation, perform requirements analysis and gathering, determine stakeholder needs, wants, and implement applications/platforms/solutions that can meet stakeholder key performance indicators (KPIs). Cahoon (2014) illustrates the space of Computer Science as a Venn diagram of AI, ML, and data mining concepts, which with the availability and utilisation of Big Data, supports and enables one another to draw out data analysis and science capabilities through the enablement of tools and technologies.

For the author's current MNC client, there is a prevalence of data – specifically Big Data due to the volume, veracity, and velocity – that requires them to leverage data analysis and science concepts, through AI, ML, and data mining –, as an integral part of what they do. Thanks to their familiarity and experience with Computer Science and carrying out data analysis on Big Data, they are able to delve deeper into business intelligence and analytics for support business' data-driven analysis and decision-making through data science in the space of optimisation of business processes, forecasting, BI, and AI/ML, as illustrated by Figure 1.1.

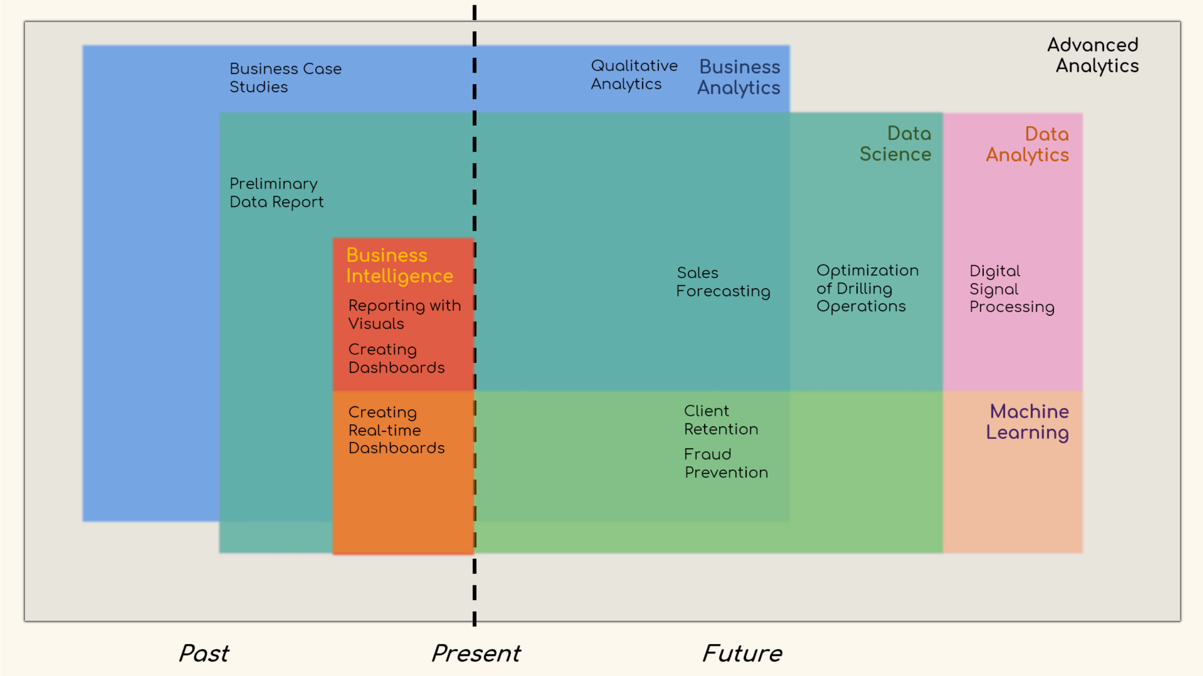

Figure 1.1. *Business Analysts and Advanced Analytics* (Sintelix, 2018).

Additionally, thanks to the author's exposure and experience with business analytics, business analysis, business intelligence, and requirements gathering and analysis, the author learnt about the data science space, motivating them to further my education and experience by getting my Master of Science (MS) degree in Information Systems and Engineering Management (ISEM) with an individualised concentration in *Analytics and Design*. To further their exposure to this field, their CPT allowed them to work directly on a data science project with their current client through practical applications of their theoretical learnings to further MS specialisation.

The *Analytics* part of this individualised concentration covers data analysis as well as predictive analysis, as well as eliciting, understanding, documenting, and expanding on the functional and non-functional requirements for analytic platforms that support those analysis capabilities – including any back-end technical components such as databases, data ingestion mechanism, and solution architecture.

The *Design* part of this individualised concentration covers the not only application and system design, but also the deliverables that the author is responsible and accountable for. The application, system, and deliverables are all meant for a human audience, necessitating that the *user*, once again, is at the centre of the design and that user experience (UX) design concepts are incorporated it – from mock-ups, process flows, visualisations, and data and document layouts. Ultimately, an integral part of a BA's job is ensuring that, as we enable change and success, the client- and user-centred view are at the heart of every deliverable and aspect of their work, which is exactly what the author embodies and continue to aim to do in every client project – they listen, deep dive, document, understand, and evaluate the needs and wants of the user at the centre of everything.

While the author has been not only able to propose a process for decision analysis (DA) on evaluating and selecting the right platform for their organisation, but also how to break into this data analytics and science space for those who are new while still ensuring that those who are already familiar in the space have the right tools to continue to succeed.

Up until three (3) years ago, the author was not familiar with the data science space, or that it even had a specific name, but as an IT BA, they do a lot of data analyses, specifically EDA, primarily through SQL and then Excel/PowerPoint to convey my findings of patterns/trends/anomalies.

In the past three (3) years alone, based on the author's research on Forrester and Gartner (Columbus, 2021) reports, there has been a rapid expansion at the corporate level of drag-and-drop data science platforms, that also support code-based data science efforts, to enable users to take the next step into actionable decisions based on their data within an organisation's ecosystem.

What the author does see is a need to for non-traditional programmers/IT resources carry out data analytics – including advanced analytics, e.g., predictive and prescriptive, through AI, DL, and ML – without the constant need for IT to be involved. They want to showcase how user enablement through a platform that emphasizes user experience can help expand data science familiarity, accessibility, utilisation, and usability. UX concepts such as product design, design thinking, UXD, and user research are what they hope to utilise as means for inspiring user analytics or for application of UXD, all skills and responsibilities that are part of being a BA.

1.3 Significance and Justification

There is a growing need for data analytics and science capabilities and skillsets (Columbus, 2021), but the question is how can those needs be met, where should one begin? With the growing data, a.k.a. Big Data, and trend of data democratisation, there is a need to have users on hand for an organisation who can learn the data, manipulate it, extract trends and patterns, and visually represent it at on an ongoing basis. The goal of that democratisation leads to increasing benefits – be they quantitative and/or qualitative. (IBM Cloud Education, 2020) (Bowley, 2017)

According to Linkedin (2017), there has been an 40% rise in positions that fall within the space of data analytics and science, with roles titled data analyst, data scientist, citizen data scientist, AI/ML analyst, etc. who's skillset includes Python, R, and SQL. Today, there are open source tools such as SQL Server, RStudio, and Jupyter notebook that support the capabilities of connecting to data, but require underlying infrastructure . That said, not all organisations are able to hire data scientists; sometimes those skills are homegrown, other times, someone who's been working with data accidentally discovers that there is a name for what they have been doing – as has been the author's personal experience from being a Business Analyst to discovering that they are also a Citizen Data Scientist.

In the marketplace, Gartner (Columbus, 2021) showcases that there has been an increase in the availability of platforms that play in the space, with a demonstrated move from the four quadrants. Focusing on just those players that have moved from the Challenger to Leader quadrant, their key capabilities are focused on serving that broad set of users and their skillsets instead of pigeonholing the skills to just those who are programming language experienced (e.g., Python, R, SQL).

Now, marrying the desire for data analysis and science with the broad skillsets available at an organisation, how does one do that? That is the process and key aspects that this paper aims to highlight, with the justification that organisations have a growing need for such a supportive platform for its employees.

1.4 Deliverables

The result of this research was to put forward a framework, utilising design thinking ideology in Figure 1.2 and the double diamond process visualised in Figures 1.3, where the central focus of an effort to evaluate and select a data analytics and science platform that supports a variety of users is based upon the human interactions required. In addition, the research aimed to highlight an approach around how to appeal to those various users in order to meet functional capabilities with the user skills and needs.

Design thinking, as shown by Figure 1.2, is a cyclical process that’s focused on carrying out research, analysing the results, ideating, prototyping, testing, implementing, and going back to the start to keep optimising.

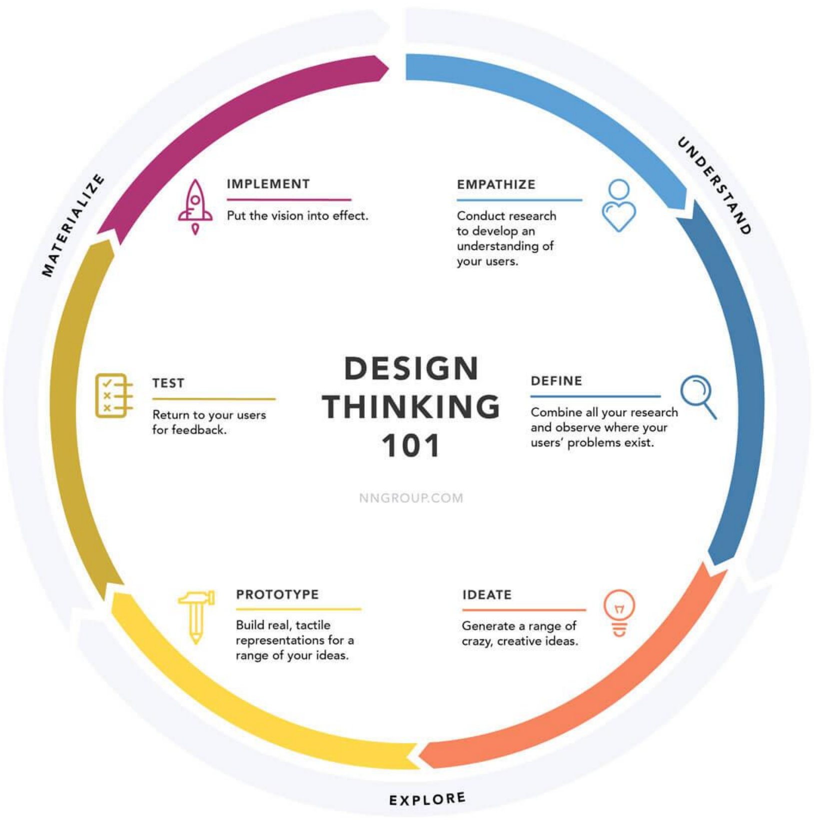

Figure 1.2. *Design Thinking 101* (Gibson, 2018).

The double diamond process, in Figure 1.3, is useful in exploring the problem space and then going into the solution space utilising the divergence-convergence model wherein ideas are gathered, analysed, understood, and then communicated.

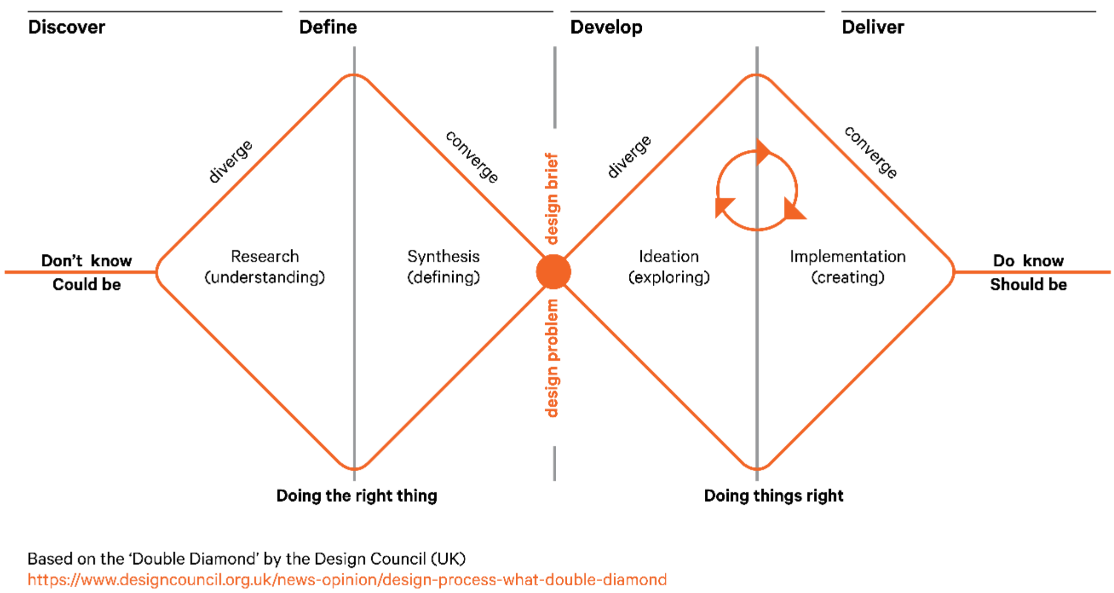

Figure 1.3. *Design Council's Double Diamond* (UK Design Council, n.d.).

Combining the design thinking and double diamond processes come documentation of the proposed templates in the appendix, from Appendix A through H. These deliverables were created and filled in via the following schedule, outlined in Figure 1.4:

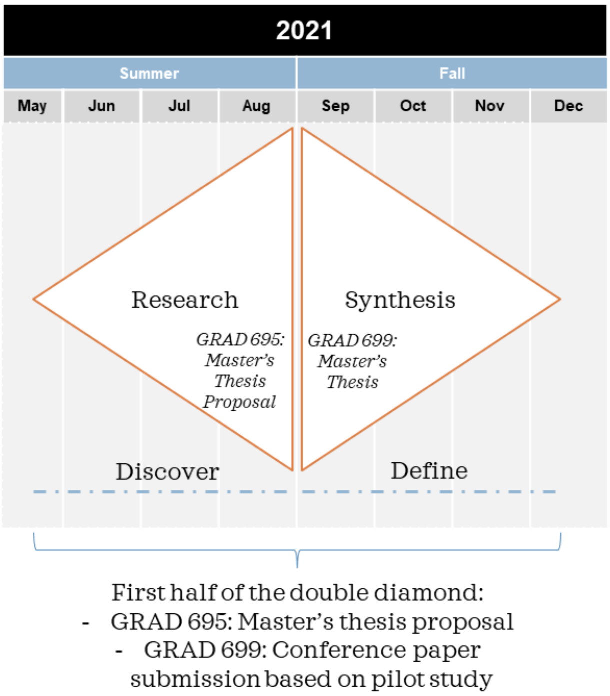

Figure 1.4. *Pilot Research Utilising the first half of the double diamond* (Aheeka, 2021).

In line with the first half of the double diamond, illustrated in Figure 1.4, the various portions of this research align up with its respective phase as follows:

- *Research phase:* The ***Error! Reference source not found.*** as well as the *Chapter 2: Literature Review: Analysis* of Related Work sections of this paper, consisting of understanding the space this problem statement is in and a literature review, aligns with the research phase as it is centred around understanding what is required to delve deeper into the research. The corresponding design thinking phase is *Empathise*.
- *Define phase:* The *Chapter 3*: Research Implementation, *Chapter 4*: Research Results, *Chapter 5: Conclusion* & Recommendations

5.0 Introduction

The goal of this research and resulting framework is its hopeful utilisation by corporations, or individual(s) at corporation(s), – regardless of the industry that they are and their role – when carrying out an evaluation for a self-service data science platform, with out-of-the-box data analysis capabilities. Individuals in roles such as Business Analysts, Citizen Data Scientists/Data Scientists, Project Managers, and/or Product Owners looking to evaluate and select a self-service data analytics/science platform would be able to leverage the framework and templates, allowing them to focus on content and the user's wants and needs instead of creating their own templates and starting from scratch. Additionally, any product vendors already in or looking to get into the DSML space have a foundation for their users that they need to augment or build their platform for, along with a set of requirements; they will also need to an understanding of a corporation's evaluation and selection framework that allows them to be better prepared for such a process.

5.1 Conclusion

The benefit of the user personas to the corporate world is the foundation definition of requirements (user wants and needs) driven by the personas experience, expertise, and skillset. Moreover, with the increased prevalence of the data analytics and science needs as well as the industry direction, this research also aims to highlight the importance in flexibility the various user personas to strengthen a company's own capabilities and success over time.

The pre-definition and categorisation of placeholder requirements in the requirements templates would allow for increased flexibility and more time for an individual, like a Business Analyst, in gap analysis and requirements gathering, which in turn would set the stage for a foundational platform needs and wants definition.

The additional templates and collected data would also allow for any other individuals wanting to re-create this study across various industries and augment the research further to not only include additional user personas but also functional and non-functional requirements as well as augmented evaluation and selecting framework for a self-service data analytics/science platform.

5.2 Limitation(s)

This pilot research was done using design thinking's prototyping style of the problem statement with ten (10) participants with the participants predominantly from the Automotive and IT industries; by expanding the number of participants to at least fifty (50), along with the industry and title representation, a broader segmentation of data categorisation can be done for personas definition and requirements elicitation.

- , and *Chapter* 5: Conclusion sections of this paper, consisting of personas, requirements, and evaluation criteria, aligns with the define phase as it is centred around defining and setting the stage for how a design problem should be addressed via a design brief or research protocol. The corresponding design thinking phase is similarly titled *Define*.

Chapter 2: Literature Review: Analysis of Related Work

2.0 Overview

The fields of data science, user self-service, and user experience are broad areas of research within their own right, and yet, are closely interconnected with one another. Central to the user self-service area is the application of user experience, as the focus is centrally on the user and the many personas that they may have, as well as the framework for enabling the data analytics and science effort by the same user set and their myriad of personas – beginner, intermediate, and advanced. To bring that full circle, the data analytics and science effort needs to be understood by those carrying out and executing it – data scientists, citizen data scientists, data analysts, and business analysts –, as well as those the process and results are presented to – unfamiliar folks, and even management and executive leadership. In other words, at the heart of data analytics and science is the user and how the information is prepared for *their* understanding, interpretation, and utilisation.

This interconnectedness arises from the rapidly growing space brought on by an increase in organisations across various industries needing the availability and adoption of data science platforms that support user-service analytics in order to enable their users of data science concepts and utilisation of their Big Data (Columbus, 2021).

The goal of the literature review is to understand the applicability of user-centred design thinking when trying to determine a process for evaluating the right data science platforms for enabling such needs, while understanding the current situation and future trends of the user self-service analytics/data science platform space. The key topics that will be analysed are data analytics & science in conjunction with user experience.

The literature review was carried out by accessing ACM Guide to Computing Literature, American Marketing Association Journal of Marketing, Academy of Management, IEEE Xplore, and Springer's Lecture Notes in Computer Science (LNCS)'s databases for peer-reviewed publications and content that contained the inclusion criteria of the relevant keyword of “user self-service” coupled with any combination of the additional keywords of “analytics”, “artificial intelligence”, “data analytics”, “data science”, “data science platform”, “machine learning”, “user experience, “UX”, and “user self-service analytics”. The subsequent search results were further restricted to publications that were in the English language and were from the year 2005 and newer to ensure the relativity of the content and applicability to a rapidly changing space; journal articles and conference proceedings were considered as acceptable results to analyse from.

The results were further limited by parsing through the title, abstract, and keywords to ensure applicability to the research topic. From a subset of thirty-four (34), four (4) met the entry criteria to be considered as a part of the thematic-centric literature review that supported a design and implementation process of data projects. From there, the following three (3) publications were selected.

In addition to peer-reviewed publications, journal articles, and conference proceedings, a Google search was also performed with the same inclusion criteria for the first hundred (100) results for the same keyword combination and permutation. The Google search yielded copious amounts of results, ultimately acting as a background for identification of applicable user self-service analytics data science platforms and related concepts for that enablement. To supplement the Google search, the Gartner 2020 and 2021 Magic Quadrants for Data Science and Machine Learning Platform as many organisation’s enterprise decisions are based off of Gartner’s “recommendations on enterprise software stack[s]” (MSV, J., 2020).

2.1 The Data Science and Machine Learning (DSML) Platform Market

From Gartner (Columbus, 2021), this user self-service analytics data analytics and science platforms fall within the data science and machine learning (DSML) platform market (Columbus, 2021) and the trend for data science and machine learning in 2021 (and beyond) proves that rapid growth in pace that are “increasingly becoming a way for companies to differentiate themselves” (Columbus, 2021). The reason for this is because it “is an adolescent market […and it has grown] by 17.5% in 2019, generating $4 billion in revenue [being t]he second-fastest-growing segment of the analytics and business intelligence (BI) software market behind modern BI platforms” (Columbus, 2021). This growth further highlights the dominance of this field across industries, and with the augmentation of capabilities from year-to-year, it ascertains the need for continued focus on ensuring that the user is at the centre of the platform.

Furthered by the segmentation of capabilities that a DSML platform provides in into the areas of “UI, augmented DSML (AutoML), MLOps, performance and scalability, hybrid and multicloud support, XAI, and cutting-edge use cases and techniques (such as deep learning, large-scale IoT, and reinforcement learning)” (Columbus, 2021), the emerging trend, driven by market demand and innovation, shines a light on what capabilities need to be part of the user self-service analytics needs. In addition to that, Gartner notes that “[t]he most innovative DSML vendors support various types of users collaborating on the same project: [D]ata engineers, expert data scientists, citizen data scientists, application developers, and machine learning specialists” (Columbus, 2021). This call-out further highlights that, again, at the core of user self-service analytics is also the personas, a.k.a. the myriad of users that users, that need to be accounted for and supported out-of-the-box.

2.2 A Brief History of Data Analytics & Science

Looking back briefly at what data analytics is "the process of analy[s]ing raw data to find trends and answer questions, the definition of data analytics captures its broad scope of the field [and …] it includes many techniques with many different goals[, such as descriptive, diagnostic, predictive, and prescriptive analytics]" (Master's in Data Science, n.d.).

Data science falls within data analytics as it is "a multidisciplinary approach to extracting actionable insights from the large and ever-increasing volumes of data [through] the scientific method, math and statistics, speciali[s]ed programming, advanced analytics, AI, and even storytelling to uncover and explain the business insights buried in data" (IBM Cloud Education, 2020). Figure 2.0 highlights the interaction of data science with Big Data, DL, ML, and AI, acting as a baseline for understanding what data science comprises of for its foundation.

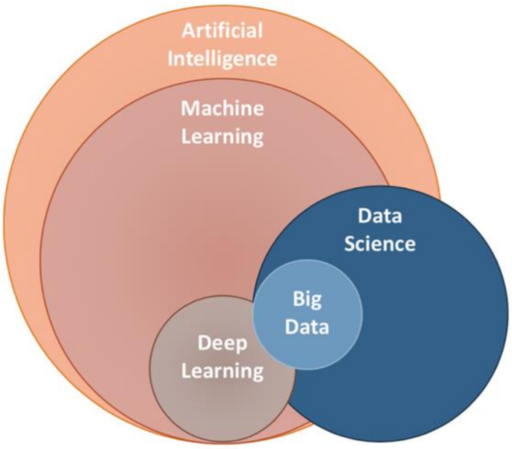

Figure 2.0. *Comparison between Data Science, Data Analysis, Big Data, Data Analytics, Data Mining, and Machine Learning* (Thakur, 2020).

Starting back in 1962, with a mathematician predicting "the effect of modern-day electronic computing on data analysis as an empirical science" (UW Data Science Team, 2017), data science has rapidly grown to become in demand, with the ability to "access, understand, and communicate the insights you get from the data analysis—are going to be extremely important" (UW Data Science Team, 2017). That growth, ultimately in 2010 started to change and give rise to consumer adoption and embracing of technology *coupled* with faster processing capabilities and growing desire to learn, understand, manipulate, gain, and present insights from the available data. As Gartner's research from 2019 onwards has shown, the data science space is growing from just those organisations that need and can afford it to players in the marketplace who offer out-of-the-box platforms that support the scalability, capabilities, and growing augmentation for users going into this space and organisations needing the enterprise technology stack to support that.

Delving deeper into what data science comprises of, Figure 2.1 maps out the iterative data science lifecycle, comprising of understanding the focus/research question, mining the data to learn what the data consists of, cleansing and preparing the data ensure good data quality, exploring the data in order to understand patterns that can be observed and device a hypothesis, feature engineering where important parameters are derived, carrying out forecasting/predictive analytics, visualising the pattern analysis results and the predictive results, and then going back and ensure that the results and findings align with the expectations and needs.

Figure 2.1. *Data Science Lifecycle* (Agarwal, 2018).

As can be seen through Figures 2.0 and 2.1, there are a multitude of applications of data science – it is not limited to just mathematics, but rather, utilises mathematics – specifically statistical – concepts and can be leveraged in the automotive, banking, finance, insurance, healthcare, manufacturing, technology, and legal fields – to name a few. In the automotive industry alone, for organisations with their manufacturing, sales, and R&D arms, some of the data science applications are around understanding how much or how many potential customers they could target successfully via regression analysis, what categorisation of their customers can occur based on their likelihood to purchase their brand through classification techniques, or even why there was a weird occurrence in performance on the factory floor through anomaly detection.

2.3 User Experience and Data Analytics & Science

At the heart of any user self-service analytics is user centred design (UCD) *and* human centred design (HCD) as the former; the two combined bridge the gap between collaboration of stakeholders and users is imperative in realising benefits and therefore must be the focus (Dollinger, et al. 2019).

However, one might ask, is all this truly worthwhile and impactful? Yes, absolutely – to enable user self-service, the user must be at the centre of the platform and usability means "[d]esigning more intuitive interfaces and workflows reduces the learning curve for lines of business and data analysts" (Columbus, 2021). By doing so, there are two direct benefits: 1. An organisation can increase the understanding of their data from a handful of power users to a larger set of users, thereby spreading knowledge, and 2. De-centralising data science and predictive analytics efforts in solely "data scientists to business analysts who prefer to iterate models on their own, often changing constraints based on market conditions" (Columbus, 2021). The flexibility not only allows real-time market change action and business decision-making that is more in line with the direct needs of the industry, but also enables growth amongst the organisation's employees by enabling knowledge sharing and increasing skillset. Additionally, "supporting multiple personas [is] a proven go-to-market strategy" (Columbus, 2021) for the DSML space and directly benefits organisations as it allows them to support the growth and cultivation of the skills in-house instead of hiring an entire team of data scientists to that work, decreasing the cost of hiring and onboarding and increasing in-house knowledge and training.

By investing in enterprise DSML platform, the underbelly is also taken care of centrally for the technology side of an organisation's house by providing out-of-the-box capabilities such as "integration and connectivity [thereby providing] platform architectures [that] are more extensible and can be tailored to an enterprise's specific needs [with the ability to…] customi[s]e machine learning models for specific industry challenges they're facing" (Columbus, 2021). This flexible, scalable, and well-integrated architecture supports and enables enterprise-wide adoption, utilisation, and expansion. Additionally, it also assists with keeping repetitive spending down due to the integration with open source software (OSS), e.g., Python (Columbus, 2021).

The immediate tangible benefits from a UCD and HCD focused approach to the data analytics and science space results in ensuring that there in fact *is* an ongoing utilisation of the platform by users in the organisation (Dollinger, et al. p. 18, 2019). When such data science platform is architected, solutioned, and implemented with the enterprise-wide utilisation and adoption in mind, with scalability in mind, it ensures that as more Business Units (BUs) and teams enter into the space, they are able to spend their time focusing on

Hand-in-hand with continued user utilisation is expansion of the user base (Dollinger, et al. p. 18, 2019) because as the proliferation of the data science occurs through its platform availability and capabilities, furthered along by an organisation's needs, so do the desires for users of varying skillsets to have and utilise such capabilities – they also grow over time to be more educated and experienced in the space.

Furthermore, with expanded and engaged users comes “modifications and additions implemented based on user feedback and suggestions” (Dollinger, et al. p. 19, 2019), leading to a specific organisation-centric data science platform that’s truly geared towards its users and their utilisation. This centrality promotes advocacy for continued widespread adoption of said platform while also ensuring that the skillset of the user base continues to grow, coming full circling with the first benefit of user utilisation.

The key findings of the literature review are centred on understanding the history behind the data analytics and science space as well as user experience with regards to user self-service analytics. There is a proven trend, that is growing at an accelerated pace, for user self-service analytics in the data analytics and science space for organisations, and that in turn requires data science platforms to be centred around *all* the users’ skillsets – be they categorised as beginners or advanced users. In line with the primary research question, the literatures reviewed set the scene for considerations for user self-service analytics in conjunction with data analytics and science is that “[t]he most innovative DSML vendors support various types of users collaborating on the same project: data engineers, expert data scientists, citizen data scientists, application developers, and machine learning specialists” (Columbus, 2021).

Chapter 3: Research Implementation

3.0 Objectives

The objective of research implementation, illustrated in Figure 3.0, was to i) define the user personas through design thinking framework, ii) develop a user self-service data science platform evaluation criteria based on the needs and wants of the personas, and iii) evaluate the usability of that framework and criteria, all while incorporating UX design concepts in order to outline interactions of the user persona with the self-service data analytics platform.

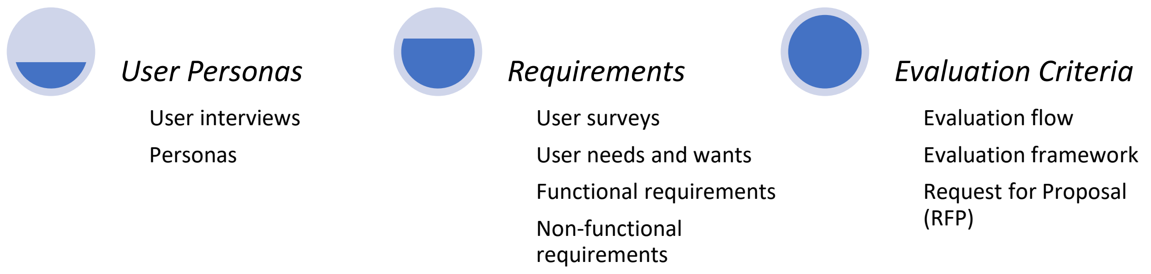

Figure 3.0. *Proposed Solution Approach for Self-Service Data Science Platform*.

In order to do so, the first step was establishing a a) research protocol, followed by putting together a b) recruitment kit – consisting of an explainer, a screener, and a consent form – for interested participants to read, understand, and sign, followed by an c) interview guide for use with the interested participants, d) carry out a fieldwork checklist for pre- and post-fieldwork, set-up a e) user persona template, define a f) requirements analysis template, and ultimately an g) evaluation framework.

The goal of defining the design thinking framework is to lay the foundation for how to corporations, or members of corporations, can utilise to select a data science platform with the core focus on the user personas and their capability requirements. To gather the user personas, user interviews and surveys must be carried out to define the skills, capabilities, and drivers of the personas. Following that, the self-service data science platform evaluation criteria (also known as requirements) will be outlined based on industry standard and persona requirements for capabilities, with metrics for those capabilities from a technical and functional perspective. Drawing those two steps together, the benefits of such an evaluation on available platform(s) will presented to the personas at organisation, regardless of their industry.

3.1 Research Execution Overview

To execute the research *A Designer's Research Manual* by Jenn and Ken Visocky O'Grady was coupled with the software development lifecycle (SDLC) process – through the lens of a Business Analyst – were utilised. The former's research-driven design framework, married with the requirements analysis process of the SDLC process led to the implementation of this proposed research approach in Figure 3.0 and documentation of the proposed templates in the appendix.

A block diagram representation of the process is available, via Figure 3.1, showcasing the preparation (inputs), data collection, data processing, and data analysis steps resulting in its associated deliverables (outputs). The preparation steps were centred around starting the research process, creating a *Research Protocol*, an *Interview Guide*, a *Survey Questionnaire*, recruiting participants, and kicking of the research participant selection steps. From there, the next step was the data collection process which consists of sending out the *Recruitment Kit* – with the *Explainer*, *Screener*, and *Consent Form* – to the selected participants so that they understand the objectives of the research and the engagement cadence, with the goal of carrying out interviews and surveys. Out of those interviews and surveys, the resulting data processing was a gathering, elicitation, and analysis of user needs & wants as well as user personas. Ultimately, as part of the data analysis process, the resulting deliverables, a.k.a. the research outputs, were categorised and defined user personas for the user self-service data analytics/science platform, associated persona-centred functional and non-functional requirements, and an evaluation flow and framework that any organisations and/or individuals can leverage.

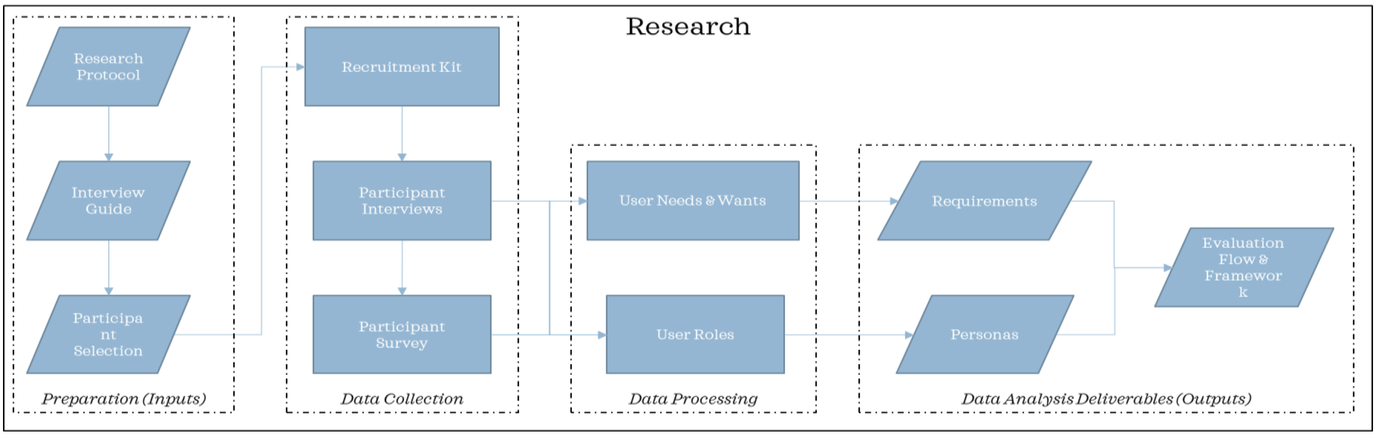

Figure 3.1. Research Implementation Block Diagram.

3.2 Research Protocol

As part of the research initiative, a *Research Protocol* in the form of a Microsoft Word Document was defined with a detailed timeline of the pilot study to be carried out on this research's proposed framework. The document contained information about the author and stakeholder(s), research project background, goals, questions, methodology, participants, schedule, budget, and placeholder interview script.

The research was carried out with graduate students and professionals in the workforce over various industries over four (4) months – from September through early December 2021 –, to determine the appropriate user personas for a self-service data analytics/science platform and then the subsequent functionalities (i.e., platform capabilities) of a platform that would serve these personas.

3.2.1 Participants

The participants the researcher was looking for this pilot research were to fall within the following characteristics:

- Current domestic or international individuals in the United States,
- Current graduate students/professionals with Internet connection,
- Active user of desktop/web applications, and
- Awareness of data analysis/science techniques.

3.3 Recruitment & Interviews

As part of the research, a recruitment period was carried out from 1st – 30th September 2021, and then ten (10) participants consented to participate, based on their understanding of the research initiative and passing the eligibility screener.

The participants who consented to taking part in this pilot research were those the researcher was familiar with and knew had some interest and/or experience in data analytics and/or science. The researched first approached thirteen (13) potential participants, with the goal of ultimately having a total of ten (10) participants consenting to participate., ranging from being graduate students to professionals. The initial recruiting was done via e-mail and text message to reach out and gauge the interest of the potential participants. From there, once interest was expressed, e-mail was utilised as a mechanism for sending the *Research Schedule*, *Recruitment Kit*, *Interview Guide*, and *Survey Questionnaire*.

3.3.1 Recruitment Kit

The *Recruitment Kit* consisted of three (3) Word Documents: 1) The *Explainer Form* that outlined an overview of what the objectives of this pilot research are and how the participant's participation would be crucial as well as utilised for analysis for the end result, 2) the *Screener Form* that is utilised to determine if the participant would be eligible in the first place for this pilot research, and allows the researcher to notify the participant of the next steps and the engagement cadence once the Kit has been sent back and reviewed, and 3) the *Consent Form* to get the participant's acknowledgement of participation and research engagement cadence, should they want to proceed and are eligible to participate in this research.

The participant received the entire *Recruitment Kit* as a Zip file with an e-mail explaining that as a first step, an overview of the objectives of this research, upcoming activities for the research, and receive consent for those who are interested in participating was being shared. Furthermore, for anyone interested, after reviewing and signing the attached *Recruitment Kit* back, there was the option of the recruiter sharing the detailed thesis proposal with the participant. The opportunity to set-up any discussions in addition to the interview and survey were also available, should the participant want it – this was to serve as a relationship building moment if the participant wanted more background into the research.

3.3.2 Survey Questionnaire

The *Survey Questionnaire*, another Word Document, was created for the participants to provide some preliminary information to research ahead of the individual interview session(s) to have a baseline of their experience in or desire to get into the data analytics and/or science space. The *Questionnaire* served as a boundary box for the participants to understand the area of focus, and to prepare the researcher better when asking questions from the *Interview Guide*.

The researcher's goal with the *Questionnaire* was to learn more about your data analysis/science experience and goals and get a baseline understanding of the participant's experience or desired experience in this space. There was a total of ten (10) questions, and these questions helped the researcher understand those areas better and there was no expectation for a right or wrong answer to any of the questions.

3.3.3 Interview Guide

The *Interview Guide*, also a Word Document, was created for the researcher to have a baseline for the questions to ask in the interview as well as to serve for the participants who have consented to the research for their preparation and background. It contained descriptive, grand tour, specific grand tour, mini tour, example, structure, and contrast questions. While the expectation for the *Guide* was that not all questions would be asked, it would serve as a foundation for questions for the researcher and individual participant to begin with.

The researcher's goal with the *Guide* was to learn more about the participant's lifestyle, skills, experience, motivators, goals, tools/technologies/services utilised/aware of, current challenges, and how they manage them as a graduate student and/or professional working with data, specifically carrying out data analysis or data science. The interview session(s) took place via Microsoft Teams, suiting a time that the individual participant was available and preferred to participate.

3.3.4 Fieldwork Checklist

Two (2) Word Document templates were created for the researcher to eventually documenting the pre- and post-fieldwork checklist. The pre-fieldwork checklist is to prepare the researcher before the interview/fieldwork and the post-fieldwork is for what has been learnt after each of the individual interviews/fieldworks are completed.

The pre-fieldwork checklist assists the researcher before an interview or fieldwork begins; it contains an equipment checklist before the interview started, any connectivity expectations, the dress code, fieldwork tools, and the environment.

A post-fieldwork checklist contains information about the equipment required in terms of what the researcher would need to do to prepare their equipment for future use, e.g., recharging; data offloading; data anonymisation; data amalgamation and aggregation.

3.4 User Personas

A Microsoft PowerPoint template was created for the researcher to eventually documenting the user personas to illustrate the “archetypical users whose goals and characteristics represent the needs of a larger group of users” (Adobe, 2019). Each persona is a maximum of one-two (1-2) slides to derive to a deeper understanding of the intended users or audience of the product – be it an application or platform/system by asking the central question of “who will be utilising the product?”. This in turn enables the development of the product to focus on the users – the who –, leading to a product that is more fulfilling for those users, meeting their needs and wants.

3.5 Requirements Analysis

Utilising Business Analysis best practices from IIBA as well as the researcher’s own professional experience in that role, a requirements template using Microsoft Excel was put together in the form of an Excel file with multiple tabs in order to capture respective information in the following tabs of the Excel file:

a) *Pain points:* To document the challenges, a.k.a. pain points, that the pilot research participant have encountered when carrying out data analysis/science processes, as well as the challenges related to the tools/technologies and platform evaluation/selection framework they utilise today.

b) *Current & Future Wants/Needs:* To document the as-is state/capabilities, a.k.a. current wants & needs, that the pilot research participant have today when carrying out data analysis/science processes via their tools/technologies and the related platform evaluation/selection framework they utilise today. It also captures the to-be state/capabilities, a.k.a. future wants & needs, that the pilot research participant would like to have when carrying out data analysis/science processes via their tools/technologies and the related platform evaluation/selection framework they would prefer to utilise.

c) *Gap Analysis:* To document the steps and/or capabilities that would take the users from the current to the future state data analysis/science processes via their tools/technologies and the related platform evaluation/selection framework they utilise today.

d) *Requirements*:

- *Functional requirements:* To document the wants and needs of the users as capabilities or functionalities that a data analysis/science platform should support for its user personas when they are interacting/utilising the platform.
- *Non-functional requirements:* To document the wants and needs of the users as operating qualities or constraints of a data analysis/science platform for compliance, management, governance, performance, and usability.

e) *Glossary:* To document terminology, acronyms, and definitions.

f) *Supporting Documentation:* For references to any other documents that are supplemental to the document,

g) *Notes:* This section was to document any information the researcher while carrying out requirements elicitation and documentation felt are applicable, e.g., assumptions, things to consider, notations, etc.

3.6 Evaluation Framework

Utilising Business Analysis and Procurement best practices as well as the researcher's own professional experience in that role, a data analytics/science evaluation and selection framework was outlined utilising Microsoft PowerPoint, with support RFP (Request for Proposal) Microsoft Excel template were put together.

The evaluation framework was to illustrate a process for any individual and/or organisation looking to evaluate and select a data analytics/science platform, but not having either a concrete/defined process or having opportunities for improvements in their existing processes that can benefit from lesson learnt from others.

3.7 Data Collection & Analysis

The data elicited from the *Survey Questionnaire* and the individual participant's interview(s) were first documented in Excel, and then the researcher categorised them into requirements – functional and non-functional – as well as defining user personas based on the participant's experiences, skills, and characteristics.

The categorisation of users' wants & needs was translated into requirements utilising industry standard nomenclature for the non-functional requirements; the functional requirements were grouped into requirement categories in the Requirements template in accordance with the researcher's professional experience as a Business Analyst.

The segmentation of the participants into user personas was based on industry terminology for various data-related roles in the data analytics/science space, utilising the participant's characteristics as a baseline for those personas; the predefined user persona template was utilised to document the information.

Chapter 4: Research Results

4.0 Overview

The research took place entirely remote as it was conducted during the COVID-19 pandemic, and therefore all the documentation for the research – both the inputs for the research as well as the outputs – were created, updated, and managed remotely utilising the *4.1 Hardware & Software* Specifications information specific in the similarly titled section below, followed by the results from participant interviews summarised in the subsequent sections: *4.2* Recruitment, *4.3* Interviews, *4.4 User* Personas, *4.5 Requirements* Analysis, and *4.6 Evaluation* Framework.

The results of the proposed research approach, illustrated as three (3) steps outlined in Figure 3.0 and detailed out in the project execution schedule in *Appendix C: Survey Questionnaire & Interview Guide*, resulted in these attached deliverables – utilising the researcher's own defined templates –, in the appendix: a) *Appendix B: Recruitment Kit – Consent, Explainer, and Screener Forms*, b) Appendix C: Survey Questionnaire & Interview Guide, c) *Appendix E: User Personas*, d) *Appendix F: Requirements Analysis & Template*, and e) Pilot Research - Requirements.xlsm.

4.1 Hardware & Software Specifications

A refurbished Dell Latitude® 7490 Business laptop with a 14-inch Full HD display (1920x1080) with anti-glare treatment running on Windows® 10 Enterprise, 64-bit OS (operating system) with Intel® Core™ i5-7300U CPU processor operating at 2.60GHz, 2.71 GHz, and 8 GB of RAM was utilised through this pilot research for all the documentation creation and management.

The laptop also had internet access and was connected to Spectrum Internet® Ultra that has up to 200 Mbps, operating at both 2.4 GHz (Gigahertz) and 5 GHz. An independently run Internet speed test by Speedtest® resulted in download speeds of 80.6 mbps (megabits per second) and upload speeds of 39.5 mpbs. The suggested download speed would be anything 6 mpbs and over to ensure that the interviews held via Microsoft Teams would not be impacted; the bigger the mbps, the better and more reliable the Internet connection, and therefore a more stable, uninterrupted interview session with each participant.

As part of the Windows® 10 Enterprise OS, the Microsoft® 365 16.0.13801.20928 64-bit version suite is available, and of there, the following applications were leveraged: i) Excel®, ii) PowerPoint®, iii) Word Document®.

4.2 Recruitment

The *Recruitment Kit* was created ahead of time, as illustrated in Figure 3.1, and utilised in this execution of the pilot research, in Figure 3.0, before the participants were selected, but after the problem statement was defined.

Of the thirteen (13) participants the researcher approached, twelve (12) initially said yes, but the final distribution of the participants was:

- *No:* One (1) participant,
- *Withdrawal:* One (1) participant, and
- *Yes:* Eleven (11) participants.

The one participant who said not was unable to support the timelines of this research with their personal and professional commitments. The participant who withdrew found that they initially misjudged their availability and their ability to contribute to this project. In both these participants' cases, the researcher was extremely understanding and grateful for the opportunity to engage with the participants, and thanked them for the time they spent trying to understand the research initiative *prior* to providing an answer.

4.3 Interviews

The *Interview Guide* and *Survey Questionnaire* were created ahead of time, as illustrated in Figure 3.0, and utilised in this execution of the pilot research, in Figure 3.1; both were sent together to each consenting participant so that they could read through and prepare, as well as decide if they would like to document their responses ahead of the interview session. Ultimately, each participant only had one (1) hour long interview, and while none of the participants were comfortable providing consent for the interviews to be recorded, they were happy to let the researcher take notes and accepting of the moments of silence where the researcher was doing so.

The actual interviews were held remotely via Microsoft Teams, with each individual interview with the participant carried out in a one-on-one virtual sessions spanning an average of one (1) hour (a.k.a. 60 minutes). The shortest interview was approximately forty (40) minutes and the longest one was approximately one (1) hour and forty-three (43) minutes.

The spread of the consenting participants who were interviewed by industry, professional title, and preferred tools for data analysis/science is visually represented and elaborated on below. Overall, given this was a pilot research with the goal to have a minimum of ten (10) participants partake in the research, there was a good representation and spread of industries and titles, allowing for the data analysis results to be reflective of cross-industry, cross-functional, multi-user data analytics/science platform wants and needs.

Diagram 4.0 visually represents the categorisation of the participant's industry and their professional title – this visual was created via a pivot table; the conclusion that can be drawn about the spread of titles by industry is that there is a good representation from a myriad of participants. There was a predominance of participants from the Automotive and IT industries. While there was an R&D Automotive industry participant, they ultimately could also be grouped into the Automotive industry as well. The only note to call out here is that there is a dominance of participants from the Automotive and IT industries, however, given that this is a pilot research, there are a total four (4) sufficient industries represented, which is a sufficient diversity. A minimum of three (3) different industries would qualify as a good diversity for a pilot research, however, if there is only one (1) industry, then at least five (5) different titles should be accounted for via the participants. The participants come from range of backgrounds and roles, thereby allowing the generalisation of data analysis/science-centric requirements gathering, persona creation, and platform evaluation framework proposal truly applicable and valid. There are nine (9) different professional titles represented, signifying a strong spread of roles, even with the predominance of certain industries over others. Moreover, by reaching a variety of roles in various industries, the data observation is that there are data analytics/science needs across the board – allowing for both generalisations to be made in parallel to specific requirements to be gathered.

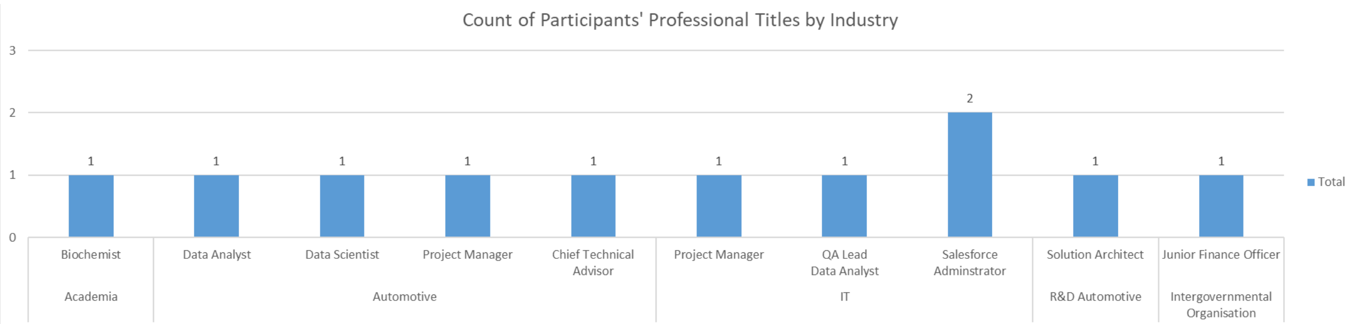

Diagram 4.0. *Spread of Participants' Professional Titles by Industry.*

Diagram 4.1 visually represents the categorisation of the participant's years of experience by industry – this visual was created via a pivot table; the conclusion that can be drawn about the least experienced participant had three (3) years of experience and the most experienced had over forty (40) years. This spread shows that at various levels of an organisation, individuals who are either new or have been there for a lot of years, there is a desire to carry out data analysis or science work.

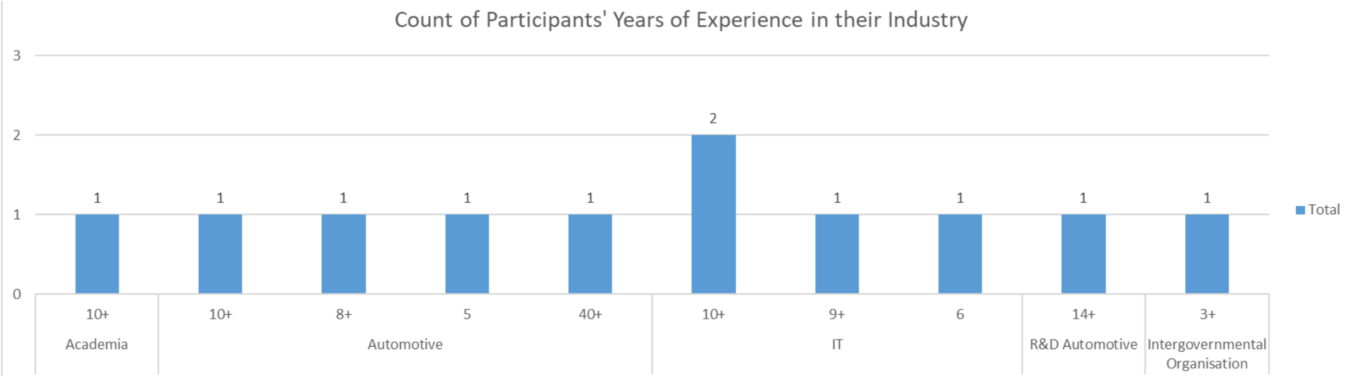

Diagram 4.1. *Spread of Participants' Years of Professional Experience by Industry.*

Diagram 4.2 showcases the spread of the primary tools utilised by the different participants; from the chart and the details provided by each participant, Microsoft Excel, PowerPoint, and outlook seemed to be the most widely utilised platforms, however, Diagram 4.3 was created to showcase the categorisation of these platforms into how they relate to and/or are utilised for data analytics/science and presentation/visualisation. However, before diving into Diagram 4.3, analysing Diagram 4.4 does provide powerful insights into what tools are being used by different professionals and the predominance – at least from these participants – of Microsoft platforms. Either from industry knowledge or searching up the tool's name the participants shared, an understanding of what these tools' primary functions can be derived and then assist in the categorisation needed for Diagram 3.0.

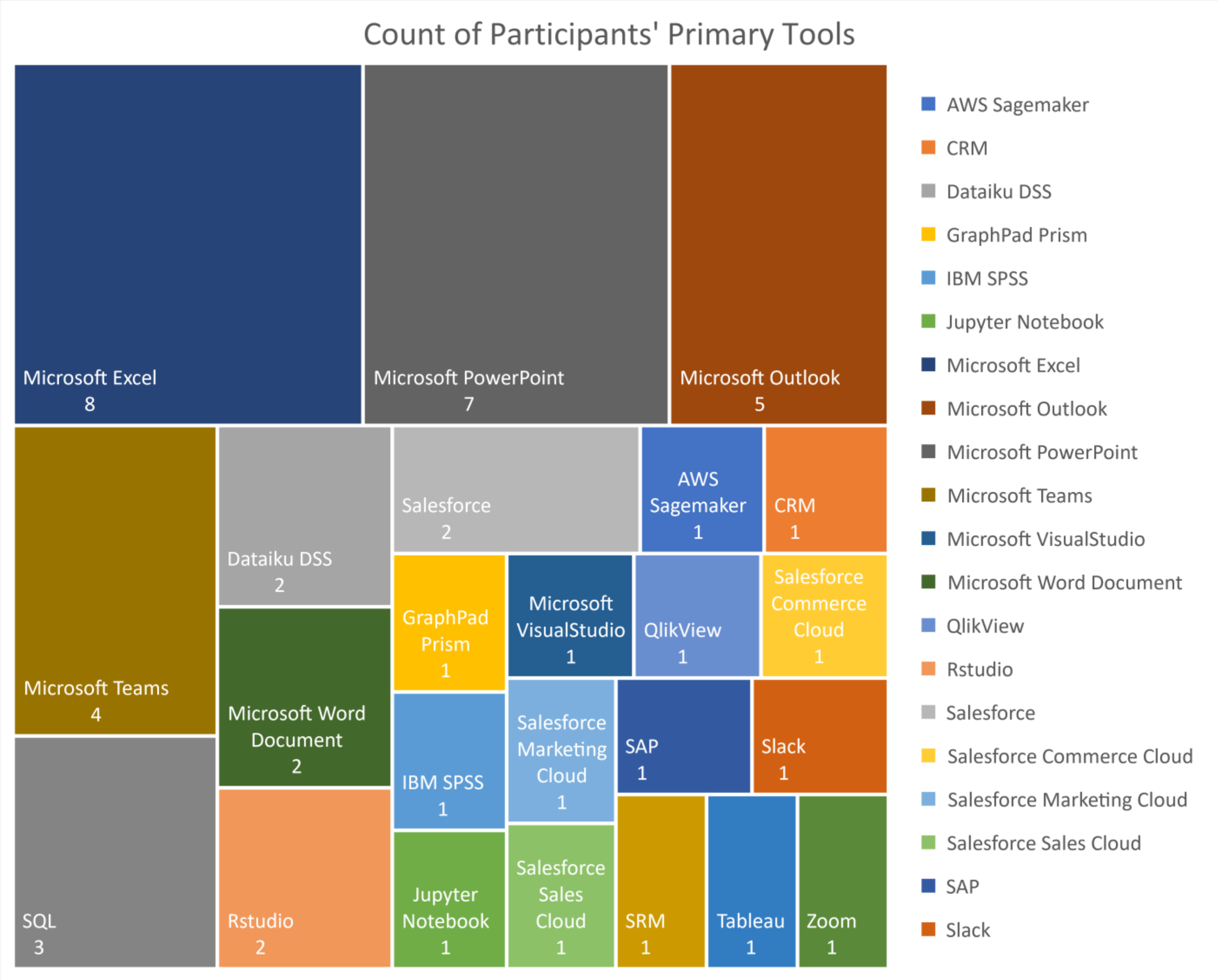

Diagram 4.2. *Spread of Participants' Primary Tools.*

Contextualising the tools provided by the participants within the research question, Diagram 4.3 was created to showcase the categories these tools would fall under when it comes to data analytics/science effort. The categories these tools fall under: i) *Communications*, ii) *Data Source*, iii) *Data Analytics/Science*, and iv) *Data Visualisation*. Here is what each category means:

- *Communications:* A platform utilised for chatting/messaging/outreaching to a colleague, a team member, a manager, etc. to relay some sort of information or to touch base.

- *Data Source:* A platform where the data (a.k.a. dataset) is made available in a readable format either for a user or another application.
- *Data Analytics/Science:* A platform utilised for viewing, analysing, preparing/massaging, exploring, and charting the dataset(s).
- *Data Visualisation:* A platform utilised for representing the dataset(s) and the results of analysis via visual images, e.g., pie charts, bar graphs, starbursts, etc.

Moreover, analysing Diagram 4.3 leads to the observation that while the majority of the tools being utilised are predominantly for one (1) purpose (a.k.a. category), there are two (2) that standout in this dataset: a) Microsoft Excel is utilised for three (3) of those four (4) categories and b) Dataiku DSS is utilised for two (2) of the four (4).

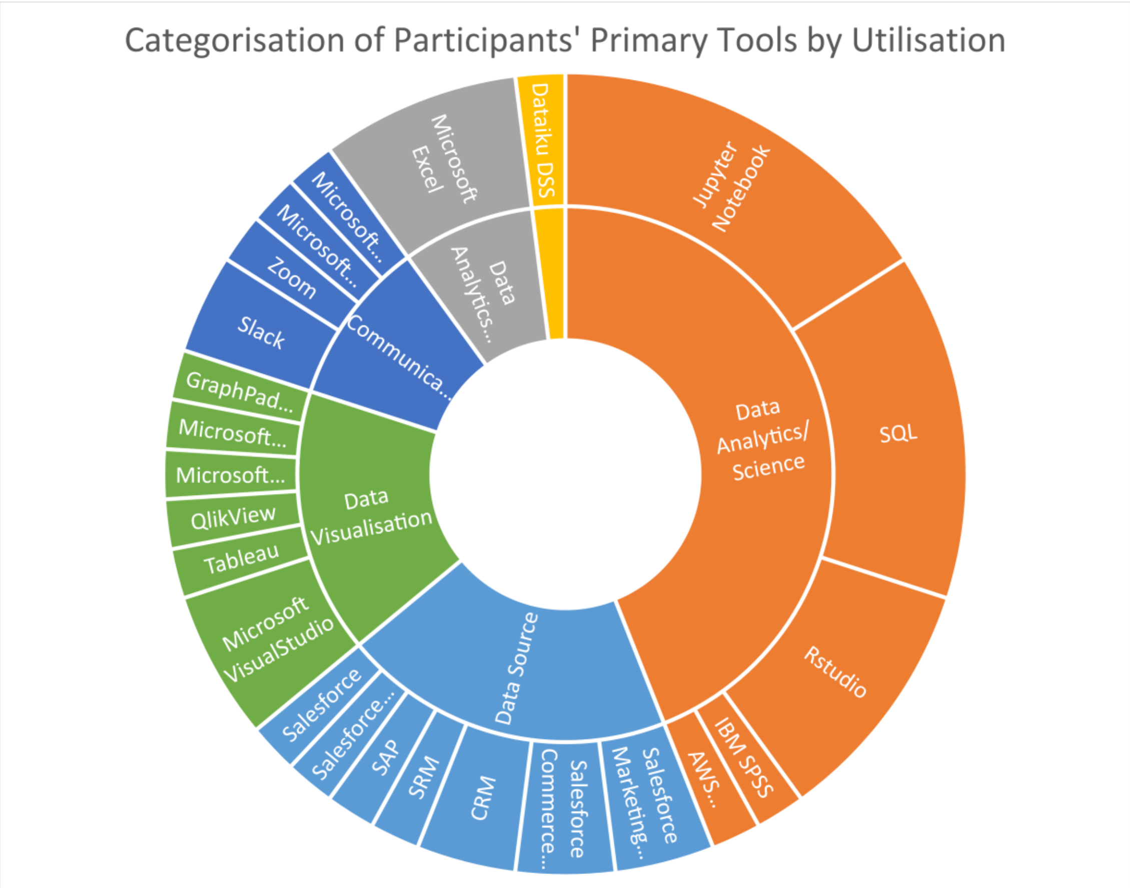

Diagram 4.3. *Categorisation of Participants' Primary Tools by Utilisation Type.*

Diagram 4.4 below allows for the mapping of which industries represented in this pilot research are solely cloud or on-premises (on-prem.), or a hybrid combination of the two, with regards to the participants own knowledge and experience. This shows that any data analytics/science platform should fall under as a SaaS (Service-as-a-Solution) or PaaS (Platform-as-a-Solution); with the former, the platform vendor can define where to host and manage the platform from, whereas in the latter, the buyer of the platform can host it as they so choose. This level of flexibility would allow buyers to still carry out data analytics/science work within their budget, resource availability, and technology stack capabilities. Six (6) of the eleven (11) participants were working predominantly with on-premises tools, three (3) were a combination or hybrid of both cloud- and on-prem.-based, and two (2) of them were cloud-based. While it looks like there is a predominance of on-prem. based tools, there is a growing shift towards cloud, with organisations supporting a hybrid approach before transitioning.

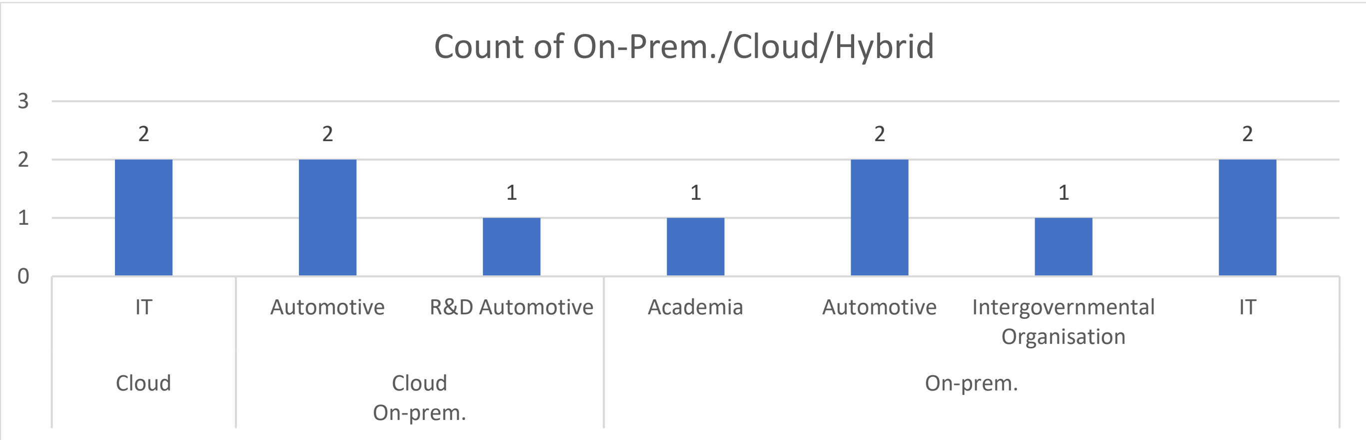

Diagram 4.4. *Prevalence of Cloud vs. Hybrid vs. On-premises Platforms by Industry.*

4.4 User Personas

Based on the eleven (11) participants' responses and background, the following personas were identified:

- *Business Analyst:* A persona focused on generating insights for others or someone who is data curious, looking to expand their understanding of data through exploration or analysis; can also be a Data Analyst.

- *Citizen Data Scientist/Data Scientist:* A persona centred on creating and/or generating predictive or prescriptive analytics, potentially within the realm of statistical analysis.
- *Data Engineer:* A persona geared more towards aligning the infrastructure/technical and data pipelines for aspects for data analysis/science.
- *Data Viewer:* A persona that is looking in on the data as a view-only user.

Having defined the applicable user personas in the data analytics/science space, the interviews were subsequently categorised resulting in four (4) of the interviewees being classified as *Citizen Data Scientists*, two (2) of the *Data Scientist*, two (2) more as Data Engineers, and one (1) as a *Business Analyst* but potentially on the path to becoming a *Citizen Data Scientist*.

Diving further into the representation of the personas by the industry, as illustrated by Diagram 4.5, the most prevalent persona in four (4) out of the five (5) industries was that of *Citizen Data Scientist* and *Data Scientist* are the most, followed by the *Data Analyst* position – identified in two (2) of the five (5) industries.

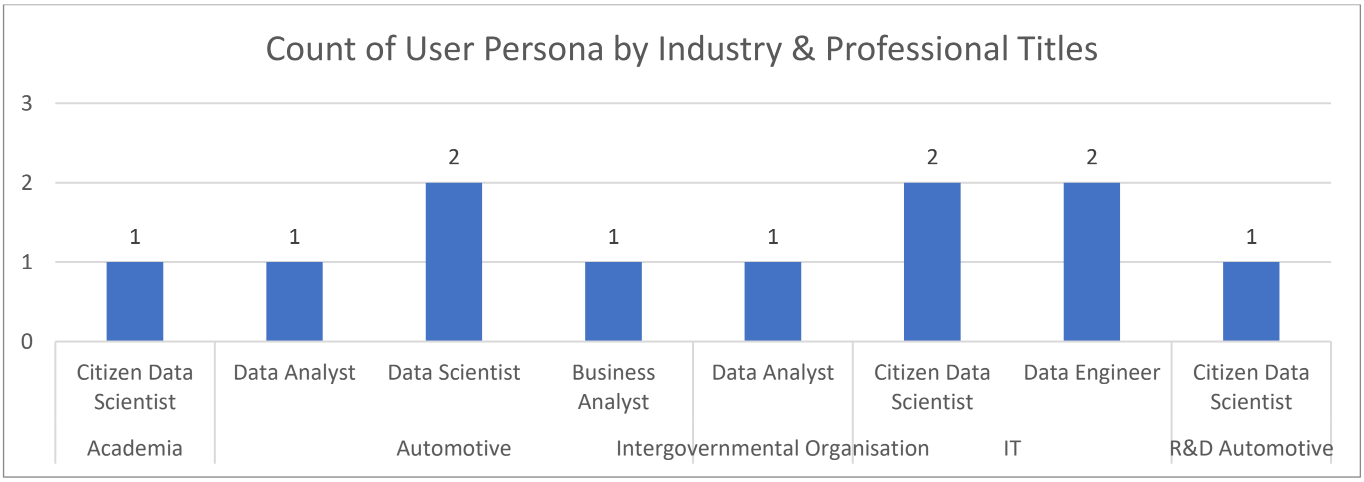

Diagram 4.5. *User Personas by Industries*.

4.5 Requirements Analysis

Utilising the participants responses in their interviews and Diagrams 4.0 through 4.5, requirements analysis was carried out by the researcher utilising Business Analyst and SDLC requirements elicitation and documentation best practices.

Based on the participant's responses on what tools they utilise for data analysis/science, what they like most of their platforms, what they wish their platforms could do, any suggestions for improvements or wishes, as well as additional drivers for that participant. After documenting the participant responses, the Business Analyst work kicked in, following the *Requirements Analysis* phase of the SDLC processes of: *Planning*, *Requirements Analysis*, *Design*, *Development*, *Testing*, *Implementation*, and *Maintenance*.

These were the resulting functional requirement categories elicited, documented, and characterised based on the participant's responses:

- *Data*: Requirements around the data sources – be it internal or external -, ingesting them – one a one-time and ongoing basis –, making it available, and storing it.
- *Data Analysis/Science:* Requirements centred around the ability for users to carry out data preparation, exploration, analysis, statistical analysis, and optimisation.
- *Data Visualisation:* Requirements to ensure the visualisation of the data analysis/science work through various charts, e.g., bar, pie, scatterplot, maps, starbursts, etc.
- *Communication & Collaboration:* Requirements to capture the communication and collaboration expectations, e.g., commenting, reviews, approvers, etc., between the various personas and additional users who may be in the periphery of the data analytics/science effort, such as subject matter experts (SMEs), reviewers, and approvers.

These were the resulting non-functional requirement categories elicited, documented, and characterised based on the participant's responses:

- *Availability/Disaster Recovery/Maintainability/Reliability/Application Monitoring/Service/Scalability:* Requirements to ensure that the availability, maintainability, reliability, monitoring, and scalability capabilities are accounted and addressed; definition of services and disaster recovery standards.
- *Data Classification/ Compliance/Management/Privacy:* Requirements to
- *Environments:* Requirement to define the
- *Scalability & Reliability:* Requirements to ensure the ability for the scalability and reliability expectations for all the applicable platforms.
- *Security & Certificates:* Requirements to ensure adherence to security protocols and regulations, along with any applicable certificates.
- *Training:* Requirements to ensure that the users are able to be trained on the platform(s) that are part of the data analytics/science ecosystem.
- *Performance:* Requirements to ensure the performance and processor speeds for the platform(s) that are part of the data analytics/science ecosystem.
- *Production Support/Help Desk:* Requirements around how the platform(s) that are part of the data analytics/science ecosystem should be supported for issues, troubleshooting, improvements, maintenance, etc.
- *Project Documentation:* Requirements to ensure there is adherence to the SDLC project methodology deliverables and documentation.
- *Users/User Groups/Roles/End Users/Stakeholders/Workflows:* Requirements around the users that need access in terms of roles and groups, as well as related workflows that are part of the different data analytics/science effort.

A visual representation of the different requirement categories – both functional and non-functional – and a count of how many requirements were gathered per category are in Diagram 4.6 below. While not all of the requirements would be applicable for all users and/or organisations, it is a foundation or template to augment further and build upon to suit their own needs and wants.

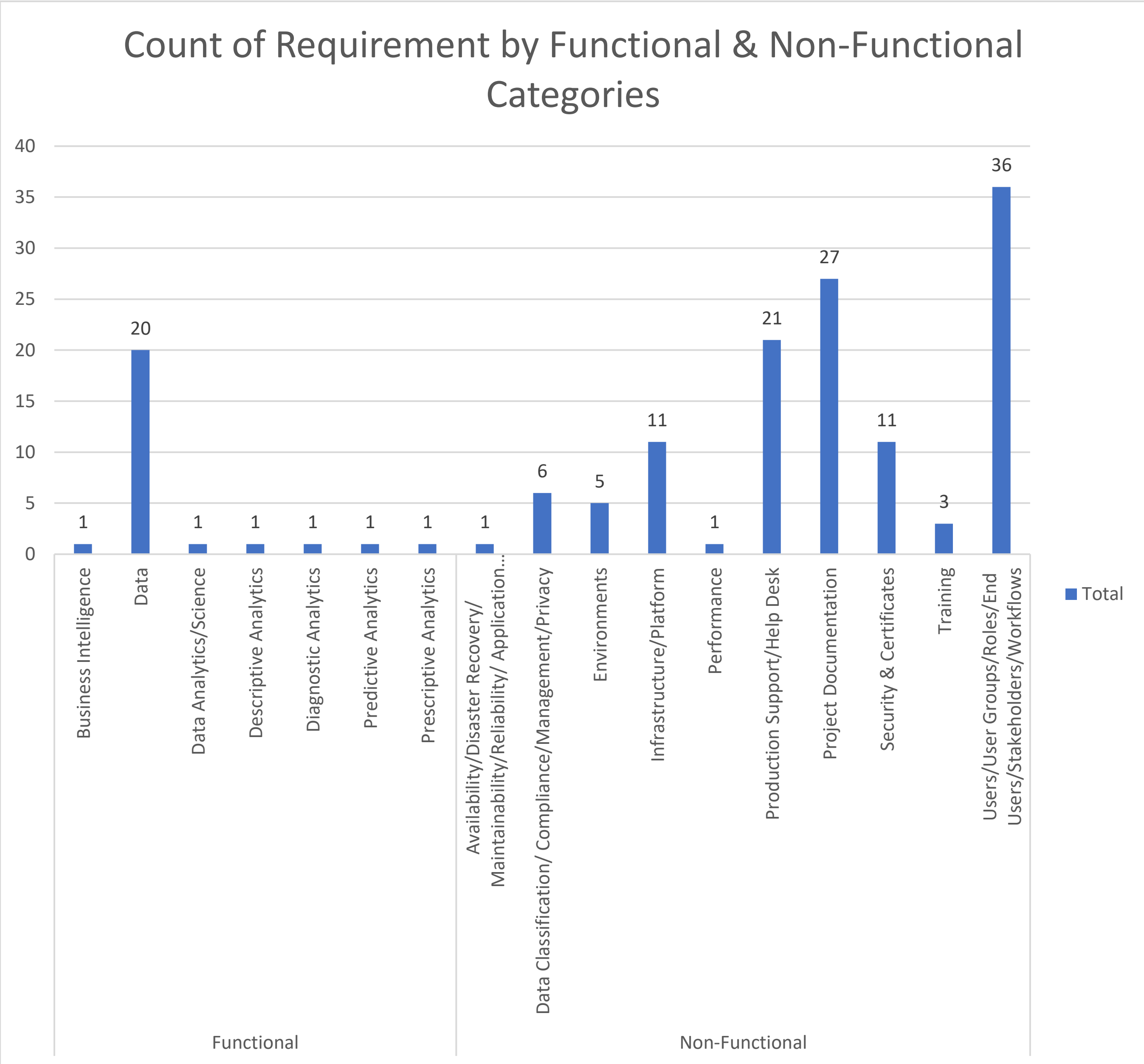

Diagram 4.6. *Count of Requirements by Categories*.

4.6 Evaluation Framework

Ultimately, the platform selection and evaluation framework is outlined in Figure 4.7 below, consisting of the following steps:

- *Ideation:* Starting with a high-level definition of what the wants and needs are.
- *Market Analysis:* After defining what requirements are applicable, along with a prioritisation associated with them, a market analysis should be carried out to understand who the players in this space are. Forrester® Wave™ for PAML (Predictive Analytics and Machine Learning) and Gartner® Magic Quadrant™ for DSML are great starting points; a search engine search can supplement the research to determine the players in this space. Typically, an organisation's analysts or their Procurement department can also help with this sort of market analysis – with the goal of determine the initial platform vendors to invite to participate in the evaluation process.
- *Vendor Wishlist & NDA (Non-Disclosure Agreement):* Once the market analysis is complete, a selection of initial platform vendors to engage with must be decided upon – those will be invited to participate in the selection process, starting with the signatures of the NDAs. For any of the vendors who do not respond with their intention to engage or do not return a signed NDA, it is a natural elimination and downselection.
- *Demo Sessions:* Once the NDAs are returned, there must be a timeboxed agenda that species the expectations of demo coverage – from how long vendor and team introductions should take to what capabilities or functionalities should be demo, aligning with the individual/organisation's requirements. Ideally at this stage, having roughly eight (8) vendors who participate would allow for a sufficient selection set to pick from.

- *Demo Score Card:* After the platform capabilities demos, the members who are the individuals/organisations looking for a data analytics/science platform need to score those demos. A scoresheet should be provided so as to contextualise and provide an audit trail for why the vendors were eliminated versus why they moved on to the next stage of the framework. While this score card will not be shared externally with the vendors, the verbal consensus can be utilised as a foundation for feedback back to the vendors who were eliminated so that they understand why they were eliminated and how they could improve. At this stage, it would be good to go from eight (8) to five (5) vendors in total so that they can go to the next stage of the process.
- *RFP (Request for Proposal):* The RFP will be created by the individuals/organisations looking for a data analytics/science platform with the intention of eliciting the financial background of the vendor, along with if the capabilities required by the individuals/organisations are available out-of-the-box, require configuration, require customisation, not yet supported but in the platform roadmap, or not supported at all. Additionally, it also serves as understanding the skillsets that the vendor could provide as part of professional services, and the associated cost & licences.
- *RFP Score Card:* Similar to the Demo Score Card, another round of scoring will occur based on the written response and Q&A session(s) held with the remaining vendors. At this stage, it would be good to go from five (5) to three (3) vendors in total so that they can go to the next stage of the process.
- *PoC/PoT (Proof of Concept/Proof of Technology) Presentations:* Much like the Demo Sessions, the PoC/PoC presentations are focused on the platform vendors diving deeper into their platform's capabilities. Here, the agenda is to

utilise data and data analytics/science use cases provided by the individuals/organisations to see how that platform would work for the buyers, and to have hands-on time in the platform environment to play around with the capabilities specific to the buyers. These presentations are also mean tot the showcase the technology stack requirements and specifications for the more technical or IT departments/teams to understand for implementation purposes.

- *DA (Decision Analysis) by Business, IT, & Procurement:* With all the prior steps, and particularly the PoC/PoT, a final decision analysis is required where it would be good to go from three (3) to the top two (2) vendors in total so that they can go to the penultimate stage of this process.
- *Procurement Negotiations:* Once the top two (2) vendors have been selected, there is also a classification of which vendor is the preferred one and which is the back-up. Then the Procurement department/team can carry out their negotiations and document finalisation – it can be SOW (Statement of Work), final agreement, and/or (PO) Purchase Order forms being ready. This can be one of the longest phases as it requires Legal and Security teams to also be involved on both sides – the buyer's and vendor's. Overall, both sides are looking ensure that they are getting the best deal for them.
- *DA – Results:* The final step of this framework results in the payment and purchase of the platform vendor the individual/organisation would like to implement as part of their work.

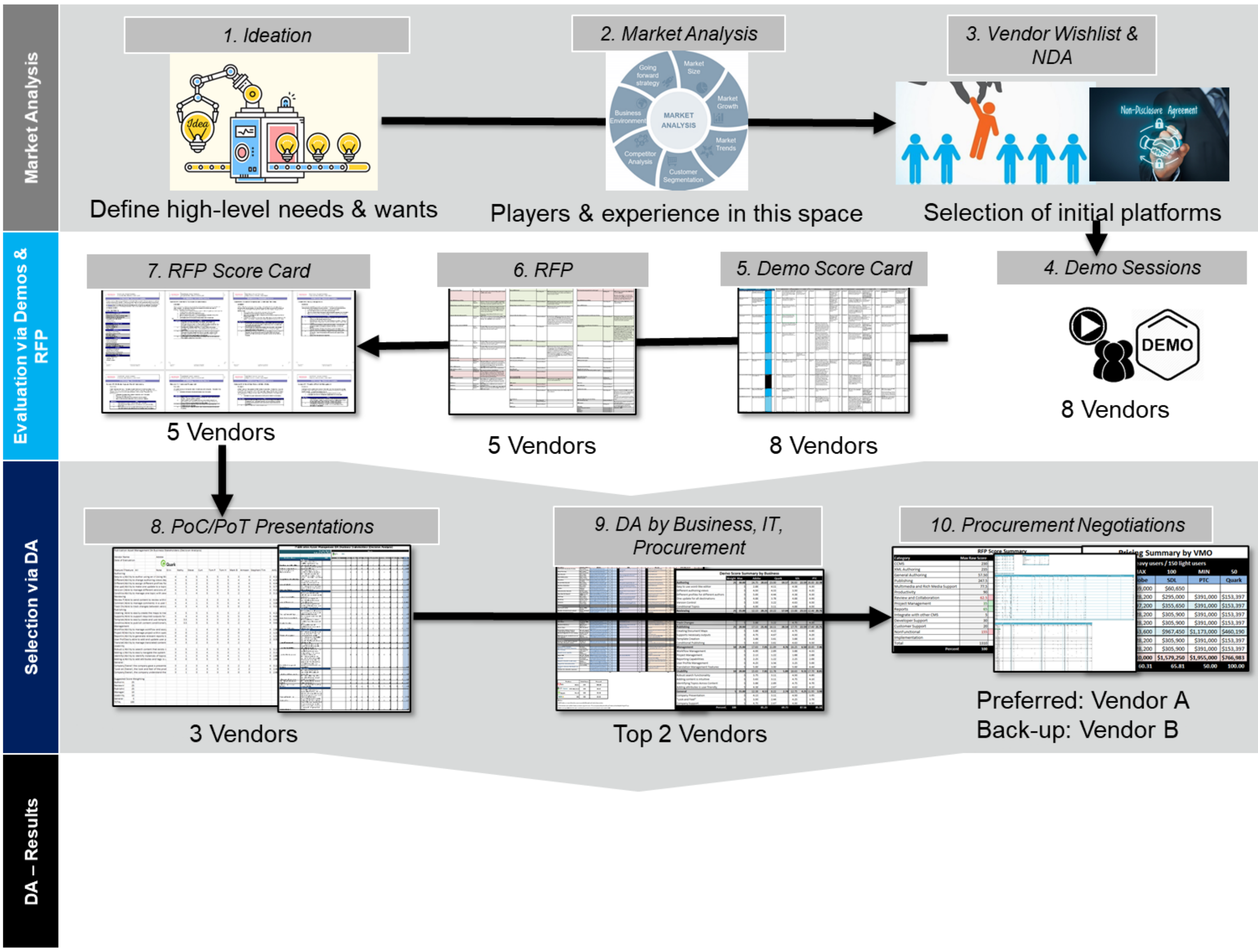

Figure 4.7. *Data Analytics/Science Platform Evaluation & Selection Framework.*

The evaluation framework outlined above in Figure 4.7 outlines the ideal process for selecting the preferred and back-up platform vendors, however for major organisations there may be standards and best practices that they will have to follow – there is a potential for incorporating in components of this framework or augmenting it as need, but that is up to the discretion of the individuals/organisations to do as they see fit.

Chapter 5: Conclusion & Recommendations

5.0 Introduction

The goal of this research and resulting framework is its hopeful utilisation by corporations, or individual(s) at corporation(s), – regardless of the industry that they are and their role – when carrying out an evaluation for a self-service data science platform, with out-of-the-box data analysis capabilities. Individuals in roles such as Business Analysts, Citizen Data Scientists/Data Scientists, Project Managers, and/or Product Owners looking to evaluate and select a self-service data analytics/science platform would be able to leverage the framework and templates, allowing them to focus on content and the user's wants and needs instead of creating their own templates and starting from scratch. Additionally, any product vendors already in or looking to get into the DSML space have a foundation for their users that they need to augment or build their platform for, along with a set of requirements; they will also need to an understanding of a corporation's evaluation and selection framework that allows them to be better prepared for such a process.

5.1 Conclusion

The benefit of the user personas to the corporate world is the foundation definition of requirements (user wants and needs) driven by the personas experience, expertise, and skillset. Moreover, with the increased prevalence of the data analytics and science needs as well as the industry direction, this research also aims to highlight the importance in flexibility the various user personas to strengthen a company's own capabilities and success over time.

The pre-definition and categorisation of placeholder requirements in the requirements templates would allow for increased flexibility and more time for an individual, like a

Business Analyst, in gap analysis and requirements gathering, which in turn would set the stage for a foundational platform needs and wants definition.

The additional templates and collected data would also allow for any other individuals wanting to re-create this study across various industries and augment the research further to not only include additional user personas but also functional and non-functional requirements as well as augmented evaluation and selecting framework for a self-service data analytics/science platform.

5.2 Limitation(s)

This pilot research was done using design thinking's prototyping style of the problem statement with ten (10) participants with the participants predominantly from the Automotive and IT industries; by expanding the number of participants to at least fifty (50), along with the industry and title representation, a broader segmentation of data categorisation can be done for personas definition and requirements elicitation.

5.3 Recommendations for Research Expansion

Any other researcher could utilise Windows® or any other OS will be sufficient, as long as there are any word processor, spreadsheet software programme, and presentation programme available for utilisation – for both creation or updates of the templates as well as filling in those templates with the resulting interview and fieldwork information. Additionally, any brand of laptop will suffice if it has the capability to run those same applications in tandem and connect to the internet – either cellular data or WiFi from an internet service provider (ISP).

Regarding the problem space, the researcher has over seven (7) years of experience utilising SDLC methodology alongside design thinking and UX design concepts, so familiarity of the terminologies as well as phases of the methodology are required to understand and expand or replicating this research.

When recruiting potential participants, it is important to reach out to more than the minimum threshold of participants required by the researcher as there may be participants who may have to drop out or other reasons that may cause them to drop out of the research. Once interested participants have been identified, send an e-mail to their preferred e-mail address with the following information – letting the participants know the engagement cadence and expectations upfront allows a trust and relationship to be built as well as a mutual understanding of the path that will be taken to achieve the end goals:

- Welcome them,
- Overview of what the research is for and how they can help,
- Explanation of each of the attached files in Zip file and the expectations for reading or filling them out,
- A schedule, and
- *Recruitment Kit* as a Zip file

The *Research Protocol* was created ahead of time, as illustrated in Figure 3.1, and utilised in this execution of the pilot research, in Figure 3.0, before the participants were selected, but after the problem statement was defined. This was a document created solely for the researcher or others who're looking to carry out similar research, with updates to include the detail scheduled once it was created. As for the remaining deliverable templates outputted as part of this project, there is flexibility in adding further categories, information, personas, and requirements. Both the *Survey Questionnaire* and *Interview Guide* can be augmented or reduced as needed. When conducting the interview, start the interview by thanking the participant for their time, delve into what the expectations of the interview are, and then talk through the questions either documents.

A lesson learnt by having a participant initially agree to the research but then having to withdraw was that preparing and providing the participation schedule as well as the *Interview Guide* and *Survey Questionnaire* upfront may have been a more tactical approach. By packaging all the information and documentation upfront, the researcher would have allowed concrete timelines to be understood by the participants and allowed that as a baseline for judging whether they could take part or not. While a general timeline had been provided in the *Recruitment Kit*, it's best to create the specific schedule closer to the actual participant selection.

Personas are also extremely suitable for communicating the findings of any research as well, particularly this research as it enables the documentation of the archetypal users who will be the users of a data analytics/science platform. For any individual/organisation looking to evaluate and select such a platform, it supports the SDLC methodology for gathering use cases, requirements, and eventually implementing a platform for their users. Additionally, for any platform vendor who wants to either get into or expand on this space, the personas enable the definition of who the users of this platform are, what their needs and wants are, as well as the out-of-the-box capabilities that the platform should support.

The information gathered via the use cases and requirements are only meant to serve as a foundation and template for any other individual looking to get into or expand in the data analytics/science space. Use cases and requirements can be adjusted as necessary – removing those that are not applicable and/or adding others that may be more specific to an individual's needs and wants. Definition of the expected production support SLA (Service Level Agreement) expectations is important as that directly impacts the users of the platform and the service availability – a demonstrated need from the participants responses. A concrete definition of what those SLAs by tiers would be a next step. Additionally, it is important to have the right stakeholders present when eliciting requirements and market analysis, competitor analysis, SWOT, PESTLE, amongst others are crucial mechanisms for determining the current platform vendors in this space. Finally, the RFP template can be utilised by any individual and/or organisation who wants to start this process of inviting platform vendors to bid in a proposal for a data analytics/science platform that would suit and fit the individual and/or organisation’s requirements. Furthermore, collaboration, data accessibility, flexibility in carrying out EDA and data science steps is crucial, usability, user experience, and is meant to be suitable for various types of personas – from novice data enthusiasts to data analysts and data scientists. Ultimately, based on the design thinking and UX design application to identifying how to evaluate a user self-service data analytics/science platform, the conclusion is that all these functionalities absolutely must be accounted for as part of the platform’s arsenal in order not only effective but also viable for the user personas.

Bibliography

Adobe. (2019, December 17). *Putting Personas to Work in UX Design: What They Are and Why They're Important*. (Adobe, Producer) Retrieved September 24, 2021, from Adobe XD Ideas: https://xd.adobe.com/ideas/process/user-research/putting-personas-to-work-in-ux-design/

Agarwal, S. (2018, February 09). *Understanding the Data Science Lifecycle*. Retrieved May 30, 2020, from Sudeep.Co: http://sudeep.co/data-science/Understanding-the-Data-Science-Lifecycle/

Aheeka. (2020). *Scaled-down data science steps.* Harrisburg University, Los Angeles, California, U.S.A.

Aheeka. (2021). *Design Thinking Framework: Self-Service Data Science Platform - Design Thinking Framework for Platform Evaluation.* Harrisburg University, Los Angeles, California, U.S.A.

Aheeka. (2021, August 29). Research Framework & Deliverables. Los Angeles, CA, U.S.A.

Ahmed, B., Dannhauser, T., & Philip, N. (2018, September 19-21). A Lean Design Thinking Methodology (LDTM) for Machine Learning and Modern Data Projects. *10th Computer Science and Electronic Engineering (CEEC)*, 11-14. doi:10.1109/CEEC.2018.8674234

Bazzano, A. N., Jane, M., Hicks, E., Maille, F., & Murphy, L. (2017, November 1). Human-centred design in global health: A scoping review of applications and contexts. *PLoS ONE*, 24. doi:https://doi.org/10.1371/journal.pone.0186744

Bell, S. J. (2009, January/February). Design Thinking. *American Libraries, 39*(1 & 2), 44-49.

Boukhelifa, N., Bezerianos, A., & Lutton, E. (2018, June 8). Evaluation of Interactive Machine Learning Systems. *Cham Springer*(Human–Computer Interaction Series), 341-360. doi:https://doi.org/10.1007/978-3-319-90403-0_17

Bowley, R. (2017, December 7). *The Fastest-Growing Jobs in the U.S. Based on LinkedIn Data*. Retrieved June 27, 2021, from LinkedIn Blog: https://blog.linkedin.com/2017/december/7/the-fastest-growing-jobs-in-the-u-s-based-on-linkedin-data

Brown, T. (2008, June). Design Thinking. *Harvard Business Review, 86*(6), 84-92. Retrieved June 15, 2021, from https://readings.design/PDF/Tim%20Brown,%20Design%20Thinking.pdf

Cahoon, J. (2014, September 14). *A.I., Machine Learning, Data Mining, and Big Data - How Do They Apply to Trading?* Retrieved May 30, 2021, from Inovance - Financial Technologies: https://inovancetech.com/buzzwords.html

Calzon, B. (2021, March 25). *Your Modern Business Guide To Data Analysis Methods And Techniques*. Retrieved May 30, 2021, from The datapine Blog - News, insights and Advice for Getting your Data in Shape: https://www.datapine.com/blog/data-analysis-methods-and-techniques/

Columbus, L. (2021, February 20). *Gartner's 2021 Magic Quadrant cites 'glut of innovation' in data science and ML*. Retrieved June 19, 2021, from PYE.AI: https://www.pye.ai/2021/03/14/gartners-2021-magic-quadrant-cites-glut-of-innovation-in-data-science-and-ml/

DataIku. (n.d.). *Predictive Maintenance Use Case*. Retrieved from Dataiku: https://dataiku.teachable.com/p/use-case-predictive-maintenance

Design Council. (2020). The Double Diamond: 15 years on. *Design Council*. Retrieved May 19, 2021, from https://www.designcouncil.org.uk/news-opinion/double-diamond-15-years

Dorst, K. (2011, August 31). The core of "design thinking" and its application. *Design Studies, 32*(6), 521-532. doi:https://doi.org/10.1016/j.destud.2011.07.006

Gibson, S. (2018, December 2). *What Is Design Thinking, Really? (What Practitioners Say)*. Retrieved from Nielsen Norman Group - World Leaders in Research-Based User Experience: https://www.nngroup.com/articles/design-thinking-practitioners-say/

Hannah, J. (2019, October 2). *What Is A User Interface, And What Are The Elements That Comprise one?* Retrieved May 29, 2021, from Career Foundry: https://careerfoundry.com/en/blog/ui-design/what-is-a-user-interface/

IBM. (n.d.). *Business Analytics*. Retrieved May 30, 2021, from IBM: https://www.ibm.com/analytics/business-analytics

IBM Cloud Education. (2020, June 3). *Artificial Intelligence (AI)*. Retrieved May 28, 2021, from IBM: https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

IBM Cloud Education. (2020, May 1). *Deep Learning*. Retrieved May 30, 2021, from IBM: https://www.ibm.com/cloud/learn/deep-learning

IBM Cloud Education. (2020, July 15). *Machine Learning*. Retrieved May 30, 2021, from IBM: https://www.ibm.com/cloud/learn/machine-learning

IBM Cloud Education. (2020, May 15). *What is Data Science*. Retrieved May 30, 2021, from IBM: https://www.ibm.com/cloud/learn/data-science-introduction#toc-the-data-s-HmAl4i7a.

IDEO. (n.d.). *IDEO Design Thinking*. Retrieved May 30, 2021, from IDEO: https://designthinking.ideo.com/

IIBA. (n.d.). *Individual Business Analysis Membership - Am I a Business Analyst?* Retrieved May 19, 2021, from IIBA: https://www.iiba.org/business-analysis-membership/individual/

Ji, H., Yun, Y., Lee, S., Kim, K., & Lim, H. (2017, July 1). An adaptable UI/UX considering user's cognitive and behavior information in distributed environment. *Sprinker Link*, 1045–1058. doi:https://doi.org/10.1007/s10586-017-0999-9

Master's in Data Science. (n.d.). *What is Data Analytics?* Retrieved May 29, 2021, from Master's in Data Science: https://www.mastersindatascience.org/learning/what-is-data-analytics/

Mieke van der Bijl-Brouwer, K. D. (2017). Advancing the strategic impact of human-centred design. *Design Studies, 53*, 23. doi:https://doi.org/10.1016/j.destud.2017.06.003.

Oracle. (n.d.). *What is Big Data?* Retrieved May 27, 2021, from Oracle: Big Data: https://www.oracle.com/big-data/what-is-big-data/

Pascal, A., Thomas, C., & Romme, A. G. (2012, January 20). Developing a Human-centred and Science-based Approach to Design: The Knowledge Management Platform Project. *British Journal of Management, 24*(2), 264-280. doi:https://doi.org/10.1111/j.1467-8551.2011.00802.x

Patil, D. (2011). *Building Data Science Teams.* Sebastopol, California, United States of America: O'Reilly.

Patton, M. Q. (2015). *Qualitative research and evaluation methods.* Saint Paul, Minnesota, U.S.A.: SAGE.

Pérez, D. (2018, February 19). *Artificial intelligence vs data science*. Retrieved May 30, 2020, from Data Science in Aviation: https://datascience.aero/artificial-intelligence-vs-data-science/

Provost, F., & Fawcett, T. (2013). *Data Science for Business: [What you Need to Know about Data Mining and Data-Analytic Thinking].* Sebastopol, California, U.S.A.: O'Reilly.

Schneider, J. (2015, February 3). *The Double Diamond: Strategy + Execution of the Right Solution*. Retrieved June 17, 2021, from Thoughtworks: https://www.thoughtworks.com/insights/blog/double-diamond

Sharp, H., Rogers, Y., & Preece, J. (2019). *Interaction design: beyond human-computer interaction* (5th ed.). Indianapolis, IN, U.S.A.: John Wiley & Sons, Inc.

Sintelix. (2018, June 18). *Machine Learning Vs Data Science Vs Analytics*. Retrieved May 30, 2021, from Sintelix: https://sintelix.com/blog/machine-learning-v-data-science-v-analytics/

Swanson, S. (2011). *Self-Service Analytics.* Sebastopol, California, United States of America: O'Reilly.

Tableau. (n.d.). *What is business intelligence? Your guide to BI and why it matters*. Retrieved May 28, 2021, from Tableau: https://www.tableau.com/learn/articles/business-intelligence

Thakur, N. (2020, April 16). *The differences between Data Science, Artificial Intelligence, Machine Learning, and Deep Learning*. Retrieved May 31, 2021, from AI.PlainEnglish: https://ai.plainenglish.io/data-science-vs-artificial-intelligence-vs-machine-learning-vs-deep-learning-50d3718d51e5

UK Design Council. (n.d.). *What is the framework for innovation? Design Council's evolved Double Diamond*. Retrieved June 12, 2021, from Design Council: https://www.designcouncil.org.uk/news-opinion/what-framework-innovation-design-councils-evolved-double-diamond

Usability.gov. (n.d.). *User Experience Basics*. Retrieved May 28, 2021, from Usability: https://www.usability.gov/what-and-why/user-experience.html

van Rijmenam, M. (2019, May 23). *7 Steps to Machine Learning: How to Prepare for an Automated Future*. Retrieved from Medium: https://medium.com/dataseries/7-steps-to-machine-learning-how-to-prepare-for-an-automated-future-78c7918cb35d

Yin, R. K. (2017). *Case Study Research Design and Methods.* Thousand Oaks, California, U.S.A.: SAGE.

Appendix

Appendix A: Research Protocol

GRAD%20695%20-%
20Aheeka%20-%20Re

Appendix B: Recruitment Kit – Consent, Explainer, and Screener Forms

Recruitment Kit Template.zip

Appendix C: Survey Questionnaire & Interview Guide

Research -
Interview Guide.doc

Research - Survey
Questionnaire.docx

Appendix D: Pre- and Post-Fieldwork Checklist

Pre-Fieldwork
Checklist.docx

Post-Fieldwork
Checklist.docx

Appendix E: User Personas & Templates

Pilot Research -
User Personas.pptx

Appendix F: Requirements Analysis & Template

Pilot Research -
Requirements.xlsm