

UNIVERSITY OF VAASA
SCHOOL OF TECHNOLOGY AND INNOVATION
INFORMATION SYSTEMS

Antti Kinnunen

DESIGN PRINCIPLES FOR A BIG DATA PLATFORM
A Value Conscious Exploration

Master's Thesis in
Information Systems

Master's Programme in Information Systems

VAASA 2019

TABLE OF CONTENTS

LIST OF FIGURES	5
LIST OF TABLES	5
1 INTRODUCTION	9
1.1 Objectives and limitations	11
1.2 Structure of the thesis	12
2 BIG DATA AND HADOOP PLATFORM	13
2.1 Big Data	13
2.1.1 Effects of Big Data in an organizational context	15
2.1.2 Knowledge Discovery from data	18
2.2 Hadoop Big Data platform	24
2.2.1 Distributed file system	26
2.2.2 Resource manager layer	28
2.2.3 Application layer	29
3 VALUE SENSITIVE DESIGN	33
3.1 Investigation types in Value Sensitive Design	34
3.1.1 Conceptual Investigations	35
3.1.2 Empirical investigations	36
3.1.3 Technological investigations	37
3.2 Critique of Value Sensitive Design	37
4 DESIGN SCIENCE AND RESEARCH PROCESS	39
4.1 Design science	39
4.2 Research process	42
5 DESIGN AND DEVELOPMENT	44

5.1 Smart Energy System Research Platform -project	44
5.2 Phase I: First technological investigation	46
5.2.1 The building of the first prototype	46
5.2.2 Prototype demonstration	48
5.3 Phase II: Second empirical investigation	49
5.3.1 Stakeholder tokens method	49
5.3.2 Stakeholder identification	50
5.4 Phase III: Second technological investigation	53
5.4.1 Securing of system resources	53
5.4.2 Design of the second prototype	54
5.5 Phase IV: First conceptual investigation	56
5.5.1 Identification of initial key values and value conflicts	57
5.5.2 Design of the interviews	57
5.6 Phase V: Third empirical investigation	63
5.6.1 Conduction of the interviews	63
5.6.2 Interview results	65
5.6.3 Harms related to stakeholders	66
5.6.4 Benefits related to stakeholders	71
5.6.5 Quantitative value prioritization by stakeholders	73
5.7 Phase VI: Fourth empirical investigation	79
5.7.1 Workshop	80
5.7.2 Workshop results	82
5.8 Phase VII: Second conceptual investigation	83
5.8.1 Value mapping	83
5.8.2 Identification and investigation of final values	87
5.8.3 Value conflict identification	89
5.9 Design principles	91

6	DEMONSTRATION OF DESIGN PRINCIPLES IN SESP-PROJECT	95
6.1	Alternatives and arguments for selections	95
6.2	Technical architecture documentation	98
6.3	Data-oriented architecture documentation	102
6.4	Platform future	104
6.4.1	Platform evolution and the lifecycle	104
6.4.2	Data sources	106
6.4.3	Challenges	108
6.4.4	Design goals	110
6.4.5	Possible practical steps forward	110
7	DISCUSSION	113
7.1	Related research	114
7.2	Limitations	116
7.3	Conclusions	117
	REFERENCES	119
	APPENDIX 1. Results of the stakeholder mapping session.	128
	APPENDIX 2. Survey Questions in Finnish.	129
	APPENDIX 3. Survey Questions in English.	131
	APPENDIX 4. Interview Warm-up Diagram.	133
	APPENDIX 5. Full Result Table of Theme 4.	134
	APPENDIX 6. The component stack and initial versions in prototype 2.	135
	APPENDIX 7. Component distributions in the cluster.	136
	APPENDIX 9. Workshop result by Team two	140

ABBREVIATIONS

Application Master	AM
Big Data Analytics	BDA
Big Data Analytics-as-a-Service	BDAaaS
Business Intelligence	BI
Command Line Interface	CLI
Database-as-a-Service	DBaaS
Data mining	DM
Design Science	DS
Design Science Research	DSR
Design Science Research Methodology	DSRM
Directed Acyclical Graph	DAG
Hadoop Distributed File System	HDFS
Hortonworks Data Platform	HDP
Information Systems	IS
Information Requirements Determination	IRD
Infrastructure-as-a-Service	IaaS
Knowledge Discovery from Data	KDD
Knowledge Discovery and Data Mining	KDDM
Lightweight Directory Access Protocol	LDAP
NoSQL	Not Only SQL
Oak Ridge National Laboratory	ORNL
Platform-as-a-Service	PaaS
ResourceManager	RM
Resilient Distributed Dataset	RDD
Smart Energy System Research Platform	SESP
Stakeholder Tokens	ST
System Security Services Daemon	SSSD
University of Vaasa	UVA
Yet Another Resource Negotiator	YARN
Value Sensitive Design	VSD

LIST OF FIGURES

Figure 1. Big Data Information Value Chain according to Abbasi et al. (2016: 6).	16
Figure 2. The advantage and data maturity. (MacGregor 2013: 28).	18
Figure 3. Elements of the knowledge discovery process. (Begoli & Horey 2012: 1).	20
Figure 4. Hadoop. (Adapted from Mendelevitch & al. 2017: 34; White 2015: 79).	25
Figure 5. HDFS architecture operation. (Adapted from Mendelevitch et al. 2017: 33).	27
Figure 6. Roles of knowledge in DSR. (Gregor & Hevner 2013: 344).	41
Figure 7. Research process used in this study.	43
Figure 8. Results of stakeholder analysis.	52
Figure 9. Network infrastructure design of the prototype 2.	55
Figure 10. The frame of reference for the interviews.	59
Figure 11. Highest prioritization concepts by points.	74
Figure 12. Co-operation and interestingness of results combined from sub-concepts.	76
Figure 13. Prioritization of the platform properties.	78
Figure 14. Workshop setup.	80
Figure 15. Final value interpretation.	87
Figure 16. Data platform overall architecture.	99
Figure 17. Data-centric overview of the platform.	102

LIST OF TABLES

Table 1. KDD Processes.(Ad. Kurgan & Musilek 2006: 6; Begoli & Horey 2012: 1).	22
Table 2. Node specification of Prototype 1.	47
Table 3. List of stakeholders chosen for further analysis.	52
Table 4. Initial values identification.	56
Table 5. Initial value conflicts.	57
Table 6. Survey participants.	64
Table 7. Emerged clusters from the interview analysis.	66
Table 8. Classes and categories within the Harms cluster.	67
Table 9. Classes inside the Potential Benefits cluster.	71
Table 10. Amount of highest prioritization.	75
Table 11. Values interpreted from Values cluster.	83

Table 12. Values identified in Harms cluster.	85
Table 13. Values identified in the Potential Benefits cluster.	86
Table 14. Identified value conflicts.	90
Table 15. Cost evaluation of cloud-based prototype 2.	97
Table 16. Master nodes partitioning table.	100
Table 17. Drive partitioning of slave nodes.	101
Table 18. Resource distribution of prototype 2.	102

UNIVERSITY OF VAASA**School of technology and innovation**

Author:	Antti Kinnunen
Topic of the Master's Thesis:	Design Principles for a Big Data Platform: a Value Conscious Exploration
Instructor:	Tero Vartiainen, Teemu Mäenpää
Degree:	Master of Science in Economics and Business Administration
Major:	Information Systems
Year of Entering the University:	2011
Year of Completing the Master's Thesis:	2019

Pages: 140

ABSTRACT:

Problem space covering the design of Big Data is vast and multi-faceted. First and foremost, it relates to the disturbance caused by the Big Data phenomenon, affecting both the people and the processes of organizations. These disturbances are a result of design choices made, both relating to technology and to the approaches used in the exploitation of opportunities offered by Big Data. These design choices are, in the end, based on the values of the designers and processed either consciously or unconsciously.

This problem space was explored with the methods of Design Science. The objective was to develop a continuously evolving and growing Big Data platform. To ensure the platform would be maintainable and developable during the whole life cycle, including situations that are impossible to foretell, it was hypothesized that by examining the purpose of the platform and by identifying consciously the values related to the platform, Big Data technologies, and to the actual usage in the envisioned environment, design principles could be created with integrating the identified values. These design principles would guide the development of the platform in the unpredictable situations of the future.

To discover the goals, benefits and the harms for the stakeholders created by the development and the usage of such a platform, methods of Value Sensitive Design were incorporated within the Design Science approach. These included empirical, conceptual, and technological investigations. During the technological investigations, two prototypes were built, the last of which will continue existence as the base of future development, and a cloud-based solution was briefly probed. Empirical investigations included project review of existing project documentation, organization of a workshop, employment of an empirical method to identify stakeholders, and the themed interviews of 16 stakeholders. Conceptual investigations were used in the identification of values.

Based on these investigations and literature seven general design principles of Big Data platforms were identified and their instantiations in the case project were described. Application of these principles in the project was also documented.

KEYWORDS: Big Data, Design Science, Value Sensitive Design, Design Principles

VAASAN YLIOPISTO**Tekniikan ja innovaatiojohtamisen yksikkö**

Tekijä:	Antti Kinnunen
Tutkielman nimi:	Design Principles for a Big Data Platform: a Value Conscious Exploration
Ohjaajan nimi:	Tero Vartiainen, Teemu Mäenpää
Tutkinto:	Kauppätieteiden maisteri
Pääaine:	Tietojärjestelmätiede
Opintojen aloitusvuosi:	2011
Tutkielman valmistumisvuosi:	2019
	Sivumäärä: 140

TIIVISTELMÄ:

Big Data -analytiikka-alustojen suunnittelu on monitahoinen ongelma. Siihen kytkeytyy ensisijaisesti Big Data ilmiön laajat uudet vaikutukset niin ihmisiin kuin ihmisten muodostamien organisaatioiden prosesseihinkin. Ilmiön vaikutukset perustuvat lopulta suunnittelussa – niin teknologiaan liittyvissä kuin myös tiedon löytämisessä – tehtyihin valintoihin. Nämä valinnat vuorostaan pohjautuvat tiedostamatta tai tiedostaen, suunnittelijan arvoihin.

Tätä lähestyttiin suunnittelutieteellisen tutkimuksen menetelmillä. Tavoitteena oli rakentaa jatkuvasti kehittyvä ja laajentuva Big Data –alusta. Jotta järjestelmä olisi kehitettävissä koko elinkaarensa ajan, myös tilanteissa joita ei voida tällä hetkellä ennustaa, oletuksena oli, että paneutumalla järjestelmän pohjimmaiseen tarkoitukseen sekä tunnistamalla järjestelmään, tekniikkaan sekä käyttöön liittyvät arvovalinnat ja ratkaisemalla ne tietoisesti, voidaan luoda järjestelmää koskevia pitkäkestoisia suunnitteluperiaatteita joissa arvot ovat integroituna. Nämä suunnitteluperiaatteet ohjaavat järjestelmän kehittämistä tulevaisuuden ennakoimattomissa tilanteissa.

Jotta järjestelmän tavoitteet, hyödyt ja mahdolliset haitat eri sidosryhmille voitiin löytää, käytettiin tutkimuksessa suunnittelutieteellisen rakenteen sisällä Value Sensitive Design –tutkimusmenetelmän toimintatapoja, mihin kuului teknisiä, empiirisiä ja käsitteellisiä tutkimuksia. Teknisten tutkimusten yhteydessä rakennettiin kaksi eri laitteistoalustoille perustuvaa prototyyppiä, joista viimeisin jäi käytettäväksi järjestelmäksi, sekä kokeiltiin pilvipalveluissa toimivia ratkaisuja. Empiiriset tutkimukset koostuivat case-projektin dokumentaatioiden läpikäynnistä, workshopin järjestämisestä, empiirisen metodin hyödyntämisestä sidosryhmien tunnistamisessa sekä teemahaastattelusta, johon osallistui 16 henkilöä. Käsitteellisillä tutkimuksilla tunnistettiin näiden perusteella järjestelmään liittyvät arvot.

Näiden tutkimusten ja kirjallisuuden perusteella tunnistettiin seitsemän yleistä suunnitteluperiaatetta ja niiden tähän yksittäiseen järjestelmään liittyvät käytänteet. Myös periaatteiden ja käytänteiden hyödyntäminen projektissa kuvattiin.

AVAINSANAT: massadata, suunnittelutiede, arvot huomioiva suunnittelu, suunnitteluperiaatteet

1 INTRODUCTION

Abbasi, Sarker & Chiang (2016: 5) view Big Data as a great disruptor, it will cause significant changes reflecting to both people and processes. Big Data related technology is not mature, it is emergent; hence the changes it causes are not fully understood nor the effects of it to people. Effects of such socio-technical systems are a direct result of how the artifacts consisting of that technology are built. Van den Hoven (2013: 78) sees these technical systems as a solidification of thousands of design decisions. These design decisions are the results of choices, and these choices embody the values of the designers (van den Hoven 2013: 78).

Simon (1996: 4–5; 1996: 114) views the artifacts as being designed to attain goals of the designer and to function, and therefore design itself is concerned how things ought to be; and everyone who is interested in “devising action aimed at changing existing situations into preferred ones” (Simon 1996: 111) is a designer. Critique of Simon’s seminal *Science of Design* by Huppertz (2015) mainly concerns the questions in the design process that Simon does not attend to, mainly the role of the designer and the how the decisions of “how things ought to be” (Simon 1996: 111) are actually done. Huppertz (2015: 40) asks whose “preferred situations” are we to design for?

Järvinen (2017: 4) sees that goals and purposes of such information systems, the preferred situations, might not be easily deducted as there can exist several groups of stakeholders, each with their own goals. Browne (2006) refers to this process of identification of the goals as Information Requirements Determination (IRD), also called requirements analysis and requirements engineering. According to Browne (2006: 313), this is considered widely as the key phase of the system development and also the most difficult.

Laplante (2014) describes several different paradigms and methods to handle uncertainty, evolving needs, and technology. The view is mostly functional, the feature is needed to do an action. Both Laplante (2014) and Browne (2006) see it as a process of logical decisions, non-functional requirements are mostly about value in the sense where it can be measured monetarily. They do not consider based on what ethical framework a design decision is actually made, decisions are made supposedly based on logic or discussions and compromises between different stakeholders. Traditional IRD methods can answer to the question of why a design decision is taken, but not to the root of the question Huppertz (2015) asks – whose preferred situations and on what grounds?

Pommeranz, Detweiler, Wiggers, and Jonker (2012) see the designers as partly responsible for creating socio-technical systems accounting for human values. Pommeranz et al. (2012) explored several different requirement engineering frameworks and methods with the aspect of eliciting situated values. Value Sensitive Design (VSD) was the only approach that explicitly focused on values while others (KAOS, SCRAM, Tropos, ScenIC, NFR) even though covering non-functional requirements and having some focus on concepts similar to values, lack specific methods for elicitation of situational values and provide little context and focus on value discovery (Pommeranz et al. 2012: 291).

If ethical values are considered in system design with scientific rigor, the very least contribution such paradigm does is that designers are aware of how their own values affect the design decisions they make. Values of the designer have a lasting impact on everyone who is somehow affected by the designed artifact, and would it not be for benefit of all, if they would be consciously processed, especially if processed in repeatable, transparent, and rigorous manner – in short, scientifically.

However, incorporating science with the design is a tremendous problem, as is demonstrated by the long evolution of design methods movement into design science, a development where further evolution is still ongoing (Cross 1993). Several different frameworks have been presented, such as the Theory of Design by Simon (1996), Information System Design Theory by Walls (1992), or Design Science Research Methodology (DSRM) by Peffers, Tuunanen, Rothenberger, and Chatterjee (2008). To design scientifically is a mighty ambition; to design scientifically *and* according to values, a mightier.

This is the challenge attempted to overcome in this thesis. By combining VSD and DSRM in an iterative research process, it is targeting the multiple design problems of Big Data platform by conducting scientific design process and making conscious choices regarding the identified values. Furthermore, by examining problematics inherent in the technological area of Big Data platforms, it is presumed design principles can be discovered highlighting important considerations in the design of such socio-technical artifacts.

1.1 Objectives and limitations

The inspiration for the thesis stems from a real-world case. An impactful unplanned change occurred in a large project at closing stages, and a data analysis and storage platform design had to be developed and implemented in a resource-limited scenario with a vague future road map of the system. It is known that technological sub-components, organizational environment, and connectivity of the system will be changing and evolving with time intervals and to directions that are not known. Driving factors behind the change can include among others developing technology, changes in participant organizations, new needs, new opportunities, improvement insights gained by practical experience of various uses of the BD platform, and budgetary changes. These changes will need new design decisions to be made, according to the state of the current system, technology, participants, resources and opportunities. It is simply too many potential scenarios of the future, too many possibilities to prepare for.

There exists duality in the objectives of the thesis. Firstly, there exists a need for a design of a data analysis and storage platform. Secondly, there is exists a need for planning the future of a system in a situation where the future holds very many potential scenarios, in an application area where technology and processes are not mature and rapidly evolving. The central hypothesis in the thesis is that *design principles* can be uncovered, to serve both aspects of objectives. In the initial design, these will assist, guide and help to conceptualize design decisions and to exist as a base on which to evaluate design trade-offs. In the future, when the designers and maintainers of the platform developed are faced with situations and needs that cannot be exactly predicted, these design principles will still be able to guide those design decisions, to exist as codified statements of the purpose and principles of the system.

These design principles will be partly based on the investigation of the values and interests of the stakeholders after stakeholders have been recognized. Partly, they will be based on the results of technological investigations, literature and industry best practices. Furthermore, it is possible that some values or important value-like aspects are so ingrained in the technological components or in the Big Data phenomenon, that it could be sensible for them to be included in the design principles. These created design principles will serve in guiding role in making design decisions regarding the system in the future, as it evolves due technology advances, changes in infrastructure or changes in the organizational processes in the external environment.

This can be expressed as the main research question of

“What kind of design principles represent the value conscious best practices of a Big Data platform?”

These suggested and nascent design principles are presumed to be somewhat generalizable, to provide sufficient demonstration and evaluation for the generalizability is outside the scope of the thesis. However, for the instantiations of the design principles, the design principles as used and created in relation to SESP-project, a demonstration will be provided.

1.2 Structure of the thesis

The thesis consists of seven major chapters. The first chapter is an introduction, explaining the background, motivation and objectives and limitations of the research. The second chapter consists of an exploration of most relevant problem areas surrounding the research: Big data as a phenomenon, Big Data in an organizational context, and how knowledge and insights can be discovered and refined. The third chapter consists of a deeper discussion of Value-Sensitive Design, theoretical basis and the tri-partite form of it. In the fourth chapter, design science is discussed in more detail. The fifth chapter is a description and documentation of the research process and results, in chronological order. Proposed design principles end the chapter. The sixth chapter consists of the demonstration of the situational implementation of the design principles in SESP-project. On seventh chapter discussion related to the results and the conclusions are presented. Lastly, references and various appendices are presented.

2 BIG DATA AND HADOOP PLATFORM

In this chapter the problematics relating to the design of Big Data platform arising from the concept of Big Data, the organizational effects, and goals of the BD platform as a socio-technical artifact, and from the rapidly changing and evolving technological environment are discussed.

2.1 Big Data

Big Data is hard to define simply. Most definitions found in the literature are based on describing different aspects of the phenomenon and the creation of a synthesis of them. As the name of the phenomenon implies, size or amount of data is one central aspect. Size is usually referred to as the first of three Vs – Volume. Most literature goes further than that and suggests additional defining features depending on the emphasis of the authors. Two additional defining features that are commonly agreed on, are Variety and Velocity which complete the “Three Vs”. (Emani, Cullot & Nicole 2015; Abbasi et al. 2016; Hashem, Yaqoob, Anuar & Mokhtar 2016; Wang, Xu, Fujita & Liu 2016; Acharjya & Ahmed 2016; Zhang, Ren, Liu, Xu, Guo & Liu 2017).

By **Volume** is defined the continuing and expanding storage of all types of data aspect of Big Data (Hashem et al. 2016:100). The volume of data is not measured in mere gigabytes or terabytes, but in petabytes and exabytes (Abbasi et al. 2016:4; Wang et al. 2016: 750). Having more data is better than having better models (Emani et al. 2015: 71).

By **Variety** aspect of Big Data is referred to the multitude of schemas found in the data and to the nearly limitless possible sources and contexts of data. It is common to group data based on the amount of meta-data available to structured data, semi-structured data or unstructured data. Structured data is data from relational databases with defined structure and relations. Semi-structured data has some attributes defined and may include data from weblogs, sensor-based data, spatial-temporal data, and social media feeds. Unstructured data has nearly no contextual information and can consist of text, raw video footage or audio recordings, for example. (Abbasi et al. 2016: 5; Hashem et 2016: 100; Eman et al. 2015: 72.)

In **Velocity** is included streams of data, the creation of structured records and availability for access and delivery (Emani 2015: 72). It is an important aspect of Big Data that it is not only concerned with data in rest. Data in motion creates new challenges as the insights and the patterns are moving targets (Abbasi et al. 2016: 5). Emani et al. (2015: 72) strongly emphasize that with Velocity is defined much more of the Big Data than the mere speed of the incoming data, the importance lies in the speed of the whole feedback loop, of providing actionable insights in time from the data in motion.

Definition of Big Data does not end in three Vs. In literature exists numerous additional Vs representing other characteristics of the phenomenon authors consider defining or important. *Veracity* has been suggested as the fourth V, representing accountability, availability, the extent the source can be trusted, accuracy, certainty, and precision (Abbasi et al. 2016: 5; Acharjya & Kauser 2016: 511; Wang et al. 2016: 750). Hashem et al. (2015: 100) is an example suggesting *Value* as the fourth V, representing what they deem the most important aspect of Big Data – “the process of discovering huge hidden values from large datasets with various types and rapid generation”.

It is possible to see the value promise represented as *Value* being important in generating interest in Big Data as a phenomenon, an important motivation why solutions for challenges within other aspects of Big Data are being pursued. Data refining and controlling for the validity of data as reminded by *Veracity* could also be argued as being important for actually gaining the Value. It is then reasonable to accept the definition of Big Data with Five Vs that includes both *Veracity* and *Value*, for example by Emani et al. (2015: 72) and Zhang et al. (2017: 3).

Additional suggestions of capturing important aspects of Big Data include Vision (a purpose), Verification (processed data conforms to some specifications), Complexity (it is difficult to organize and analyze Big Data because of evolving data relationships) and Immutability (collected and stored Big Data can be permanent if well managed) and were found by Oussous, Benjelloun, Lahcen and Belfkih (2017). There exist many more proposed additions in literature.

A multitude of proposed definitions for different aspects of Big Data exists as Big Data can be viewed from several distinct perspectives. Wang et al. (2016: 749) recognize the product-oriented perspective, the process-oriented perspective, the cognition oriented

perspective, and the social movement perspective, each with different definitions. Hashem et al. (2015: 100) propose the following definition based on their analysis of various definitions “*Big data is a set of technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex and of a massive scale*”. It incorporates the other perspectives Wang et al. (2016) mentioned, the exception being the social movement perspective which is referred to only weakly via “large hidden values”.

2.1.1 Effects of Big Data in an organizational context

Abbasi et al. (2016: 3) define information value chain as a “cyclical set of activities necessary to convert data into information and, subsequently, to transform information into knowledge”. They see Big Data essentially as a big disruptor and recognize three ways socio-technical systems and their operation in organizations change. Firstly, new information value chain requires different roles, processes, and technologies. Secondly, they see movement towards the fusion of technologies into “platforms” and in the knowledge-derivation phase transformation of processes into “pipelines”. Thirdly, they see a greater need of people who can refine data into information and eventually to knowledge, data scientists and analysts, in the all phases of the value chain to support self-service and real-time decision making. (Abbasi et al. 2016: 5).

The information value chain in the era of Big Data according to Abbasi et al. (2016) is illustrated in figure 1. Abbasi et al. (2016) see the value of data in the organizational context in the knowledge derived from the data, which in turn enables decision making that leads to actions. Results of actions produce more data and provide feedback data that, once refined to knowledge, can be used to base new decisions on. (Abbasi et al. 2016: 5–6).

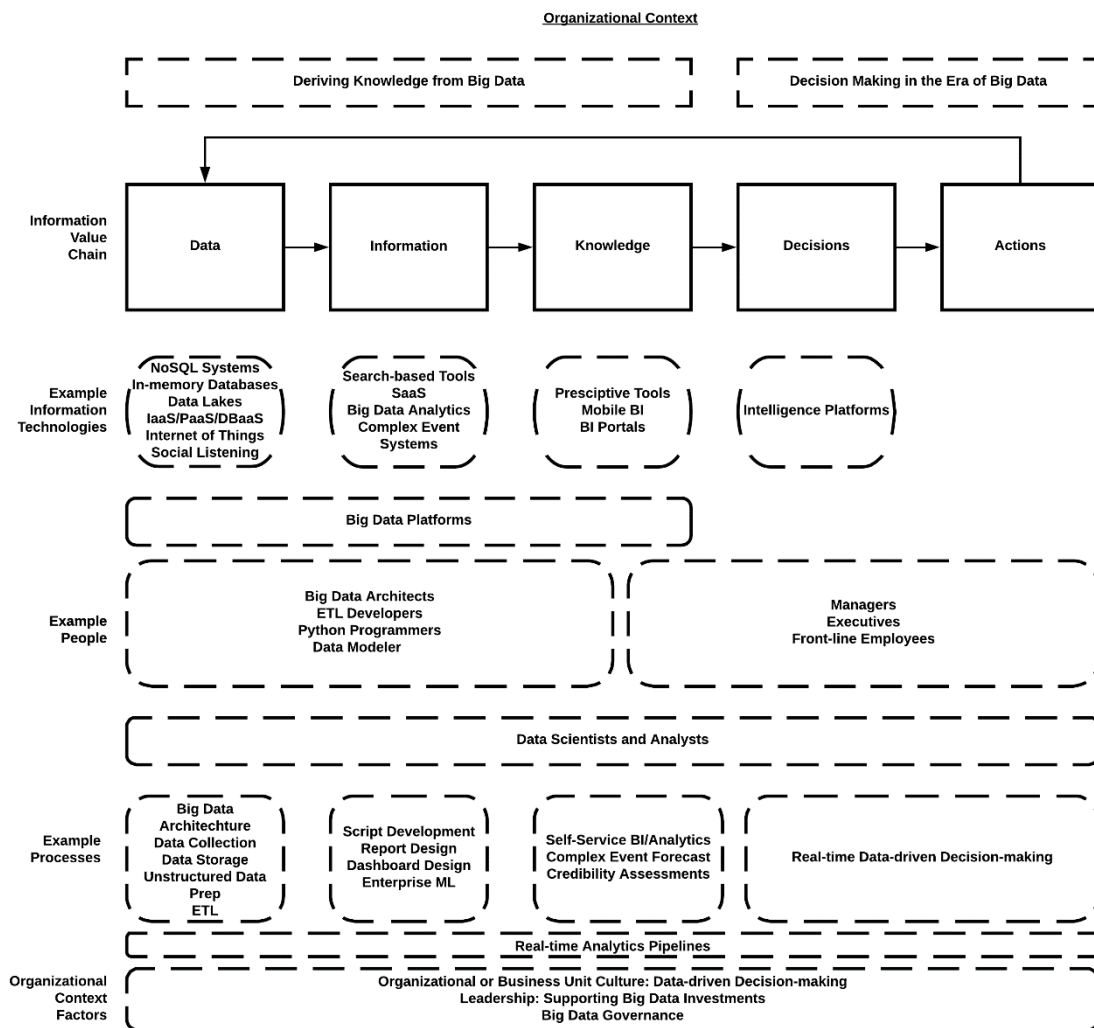


Figure 1. Big Data Information Value Chain according to Abbasi et al. (2016: 6).

Abbasi et al. (2016) suggest that previously mentioned qualities of Big Data have caused organizations to move from traditional systems of data warehouses and databases towards distributed computing and storage, to systems leveraging Hadoop or, in addition, to in-memory database solutions such as Spark to be able to gain insights and cope with rapidly incoming unstructured data with large volumes. Abbasi et al. (2016: 6) see “that the key-data management and storage questions that practitioners pose have shifted to: ‘what other internal/external data sources can we leverage’ and ‘what kind of enterprise data infrastructure do we need to support our growing needs?’”. Another technological change is a movement towards cloud-based services instead of on-premises data services such as

infrastructure-as-a-service (IaaS), platform-as-a-service (PaaS), database-as-a-service (DBaaS), and even Big Data Analytics-as-a-service (BDaaS). (Abbasi et al. 2016: 5; Damiani, Ardagna, Ceravolo & Scarabottolo 2017: 5).

There are several reasons for this shift towards cloud-based services in Big Data Analytics (BDA). Some of them that are congruent with a general shift in ICT towards cloud-based solutions including virtualized resources, parallel processing, security, data service integration with scalable storage and resulting improved efficiency in infrastructure maintenance, management and user access (Hashem et al. 2014: 99). Velocity and thus rapidly increasing Volume are aspects of Big Data that make these efficiency gains attractive for organizations (Abbasi et al. 2016: 5). Secondly, Abbasi et al. (2016: 5–7) see these changes resulting in complex Big Data architectures with multiple processes, which need a new kind of knowledge. This combined with the emerging nature and immaturity of technological components can mean that for some organizations, outsourcing and cloud-based solutions are the only reasonable way to acquire some of the human resources needed in designing, building, maintaining and using a BDA solution specific for their organizational needs.

Davenport & Patil (2012) recognize the difficulties in finding, assessing and holding on to data scientists that they define as “the people who understand how to fish out answers to important business questions from today’s tsunami of unstructured information” (Davenport & Patil 2012: 73). According to Davenport & Patil (2012: 74), a data scientist must be able to write code, have a business understanding and have the ability to find stories in the data, provide a narrative for it and to be able to communicate the narrative effectively. Abbasi et al. (2016: 6) see the data scientist working closely with analysts and management in the decisions making phase.

Effects of Big Data are comprehensive in organizational decision-making level. According to Abbasi et al. (2016: 7) Velocity of Big Data combined with the general trend toward data-driven decision making have changed how organizations both create and leverage knowledge for decision making. Like Emani et al. (2015: 72) suggested, with Velocity is not only suggested the speed of the incoming data, but also the speed of the Information Value Chain. Abbasi et al. (2016: 7) distinguish one of the biggest shift as organizations consuming analytics in real time. They see self-service business intelligence (BI) and analytics run independently by various employees, including managers and executives, a central factor in how organizations can keep up with the fast pace and

complexity of the marketplace. It makes possible agile decision making without reliance on IT or decision analyst support (Abbasi et al: 2016: 7).

2.1.2 Knowledge Discovery from data

At the root of both the Value aspect and in the value promise of Big Data is the process of refining knowledge out of raw data. This process is illustrated in figure two on a general level according to MacGregor (2013). Essentially, the challenge is that from raw data itself not much competitive value can be gained. If it is straightforward to use with simple reports, it is likely that others are also utilizing it. It has to be refined, processed, analyzed and models created to gain competitive advantage. The more processed and refined the data is and the better the models created based on the data are, the more valuable questions can be answered. Questions that can be answered transform as analytics and data maturity grows from understanding what happened, to understanding reasons and causes, onwards to prediction and finally to optimization. (MacGregor 2013).

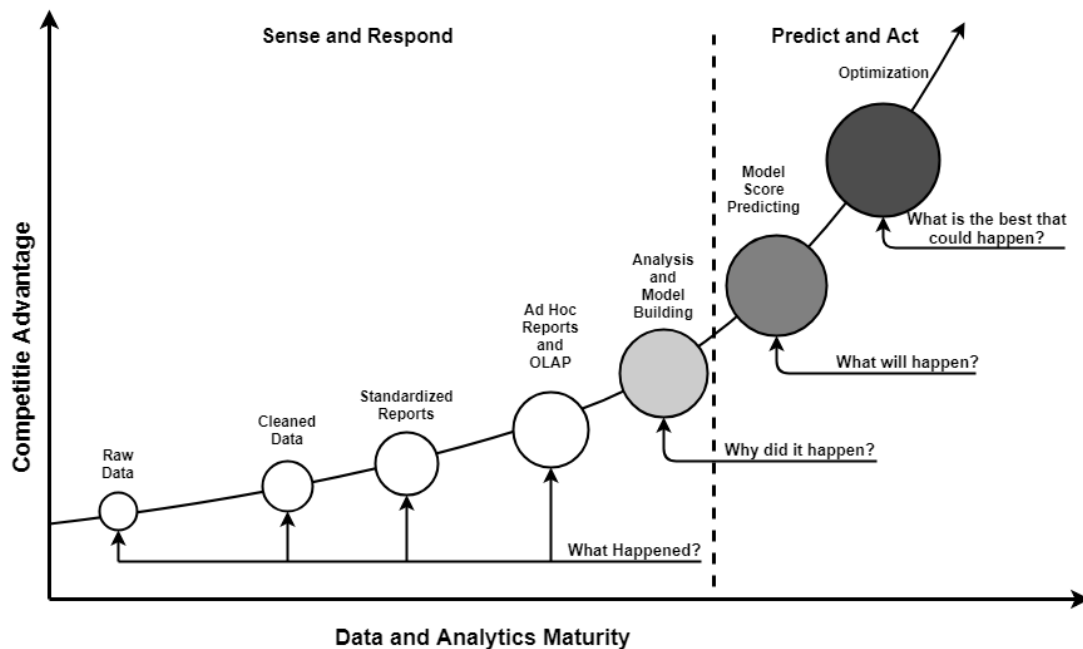


Figure 2. The advantage and data maturity. (MacGregor 2013: 28).

As time and resources are spent on analyzing the data, an investment is made. Depending on the actual process, people involved, a technological foundation in terms of the system storing the data and analytical tools, and data available an investment may lead to insight or knowledge gained. Which, in turn, may lead to competitive advantage in general, be it scientific or economic specifically.

Not all raw data is useful. Problem is deciding which raw data has value, as with large volumes of interconnected data valuable insights can be gained that are not obvious. For example, in Holland was an attempt underway to meet the EU CO₂ targets by increasing the efficiency of the electrical grid by installing smart meters in all households (Hoven 2013: 75). These smart meters recorded energy consumption every seven seconds and once measured and diligently stored into a database, provided a surprisingly good view of what was happening in the household (Hoven 2013: 75). As it was possible to even find out what movie was being watched by combining data sources and in the design the importance of the value of privacy was forgotten, eventual public concern regarding privacy rose to the level that the proposal did not pass in the Dutch upper house of the parliament (Hoven 2013: 75).

On a more general level, the process of refining data for competitive advantage can be described as discovering knowledge out of data. Knowledge as a concept differs from information. Wiig (1993: 73) defines knowledge as “knowledge consists of truths and beliefs, perspectives and concepts, judgments and expectations, methodologies and know-how”. Information, on the other hand, “consists of facts and data that are organized to describe a particular situation or condition” (Wiig 1993: 73). Information is what is gained in the earlier phase of data and analytics maturity. Knowledge is gained in the later phase. According to Wiik (1993: 73) knowledge is used to interpret information, to understand the situation and what the information means.

As the data matures through applied transformations and is processed with more and more developed analytics, slowly output becomes knowledge instead of information. There is no distinct line when that change occurs. Wiik (1993: 73) sees that as information is received by experiences and analyzation, it is gradually organized and internalized and it becomes knowledge.

Begoli & Horey (2012: 215) define Knowledge Discovery from Data (KDD) as a “set of activities designed to extract new knowledge from complex datasets”. They identify three

parts the KDD processes are comprised of. Firstly, data collection, storage, and organizational practices, secondly understanding and effective application of the modern data analytic methods (including tools) and thirdly, understanding of the problem domain and the nature, structure, and meaning of the data. This is illustrated in figure three. (Begoli & Horey 2012: 251).

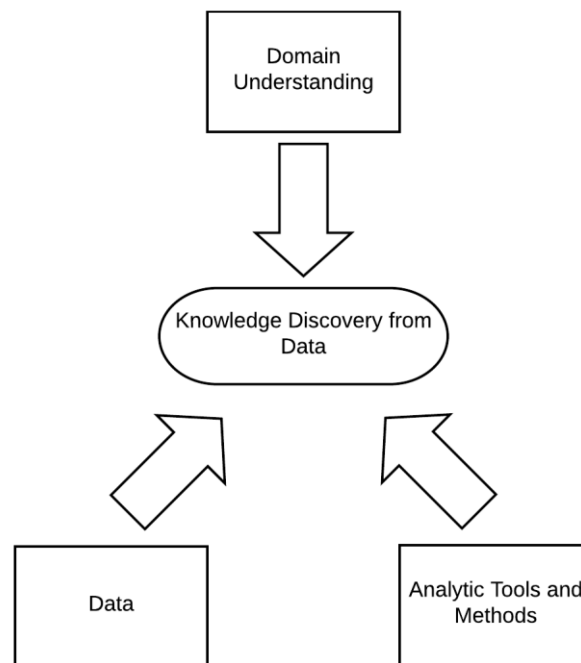


Figure 3. Elements of the knowledge discovery process. (Begoli & Horey 2012: 1).

Data Mining is a common term used when discussing this process of gaining actionable insights from data. In the definition of KDD by Begoli & Horey (2012) it is included in the more general description of “Analytic Tools and Methods”. Kurgan and Musilek (2006: 2) define Data Mining (DM) as “application, under human control, of low-level DM methods which in turn are defined as algorithms designed to analyze data, or to extract patterns in specific categories of data”. They see Knowledge Discovery (KD) as “a process that seeks new knowledge about an application domain. It consists of many steps, one of them being DM, each aimed at completion of a particular discovery task, and accomplished by the application of a discovery method” (Kurgan & Musilek 2006: 2).

Further, they define Knowledge Discovery and Data Mining (KDDM) as the KD process applied to any data source. Thus, KDD defined by Begoly & Horey (2012) is congruent with the definitions proposed by Kurgan & Musilek (2006). It is the KD process applied to complex data. In this thesis Knowledge Discovery from Data (KDD) will be used.

Kurgan & Musilek (2006) performed a survey of different KDD processes. They identified four main motivational factors for formally structuring KDD process. Firstly, the application of DM methods without an understanding of input data has the potential to lead to the discovery of knowledge without use. Validity, novelty, usefulness or understandability of the results is lacking. The main reason for defined and structured KDD process is that only by the application of such a process can result with that kind of properties be achieved. (Kurgan & Musilek 2006: 2–3).

Second identified factor raises mostly out of human cognitional limitations. Confronted with high volumes of varied data, it is hard to gain a holistic view and understanding of both data itself and the potential of the data. Kurgan & Musilek (2006) propose that this leads commonly people to rely on domain experts to gain understanding and this behavior could be attributed both to uncertainty relating to new technology and to the uncertainty of the process needed. It is their conclusion that this creates a need for both popularization and standardization of methods in this area. (Kurgan & Musilek 2006: 3).

Thirdly, structured KDD process is needed for management support. It is common for the KDD process to be part of a larger project or solution and involve co-operation of a varied number of people, departments or other actors. Without a structured process, the management of the KDD process in terms of budgeting or scheduling can be problematic. Structured KDD process also helps in communication. It makes it easier for the management and other professionals involved to get a concrete idea of what the process involves and how it proceeds. (Kurgan & Musilek 2006: 3).

Fourth and last motivational factor for formally structuring and to standardize KDD process Kurgan & Musilek (2006) identified is the need for a more unified view on existing process descriptions. This would allow the use of the emergent and constantly evolving usage of appropriate technology in solving current business cases. (Kurgan & Musilek 2006: 3).

To design and implement a Big Data analytics platform, it is essential to understand the KDD process in addition to the organizational processes involved in operations. A sample of major existing KDD processes which Kurgan & Musilek (2006) compared is presented in table one. Chosen were the most influential academic model of the time (Fayyad, Pi-atetsky-Shapiro & Smyth 1996), EU-backed industrial model called Cross-Industry Standard Process for Data Mining (CRISP-DM) developed by consortium of DaimlerChrysler and SPSS, and a generic model proposed by Kurgan & Musilek (2006) as a synthesis of all the models examined.

Table 1. A Sample of KDD Processes. (Adapted Kurgan & Musilek 2006: 6; Begoli & Horey 2012: 1).

Model		Fayyad et al. (1996)	CRISP-DM	Generic Model
Area		Academic	Industrial	N/A
No of Steps		9	6	6
Steps	Domain Understanding	1. Developing and Understanding of the Application domain	1. Business Understanding	1. Application Domain Understanding
	Analytic Tools and Methods & Data	2. Creating a Target Data Set	2. Data Understanding	2. Data understanding
		3. Data Cleaning and Preprocessing	3. Data preparation	3. Data Preparation and Identification of DM Technology
		4. Data Reduction and Projection		
		5. Choosing the DM Task		
		6. Choosing the DM Algorithm		
	7. DM	4. Modeling	4. DM	
	Knowledge Discovery	8. Interpreting Mined Patterns	5. Evaluation	5. Evaluation
		9. Consolidating Discovered Knowledge	6. Deployment	6. Knowledge Consolidation and Deployment

When comparing the steps on all three sample KDD processes, it is straightforward to see how the elements described by Begoli & Horey (2012) are present in all sample KDD processes. These elements have been fitted to the original table and to have light shading.

Begoli & Horey (2012) propose three principles for effective knowledge discovery from Big Data, based on their experiences on real-world projects at Oak Ridge National Laboratory (ORNL). ORNL works in close co-operation with different state and federal agencies on Big Data projects. ORNL receives the data, has the responsibility to analyze it with domain experts and to present the results via various avenues. Analysis techniques are not always defined, and they have to explore methods available. They also perform re-evaluations of the Big Data infrastructures and strategies of various state and federal agencies. (Begoli & Horey 2012: 215).

Three principles Begoli & Horey (2012: 216–217) propose for effective knowledge discovery from Big Data are as follows: 1) *Support a Variety of Analysis Methods* 2) *One Size Does Not Fit All* and 3) *Make Data Accessible*. All principles have subprinciples and are next presented in a summarized form.

With the first principle, *Support a Variety of Analysis Methods*, Begoli & Horey (2012: 216) mean that in KDD and in modern data science is employed a diverse group of methods from different fields, examples they mention are distributed programming, data mining, statistical analysis, machine learning, visualization, and human-computer interaction. A different set of tools and techniques are often applied in each. Different data and different kind of analysis require different kinds of expertise. For Big Data platform to enable proper analyzation of multiple kinds of data with various fields of expertise, it must support a variety of methods and environments. In ORNL following have been frequently used 1) Statistical analysis 2) Data Mining and Machine Learning and 3) Visualization and Visual Analysis. (Begoli & Horey 2012: 216).

The second principle, *One Size Does Not Fit All*, relates to the idea that a good, flexible Big Data platform offers a means for storing and processing the data at all stages of the pipeline. Their main argument is that “different types of analysis and intermediate data structures required by these (e.g. graphs for social network analysis) call for specialized data management systems” (Begoli & Horey 2012: 216). They have support for their view

that the era of generalized databases is over. They have three specific recommendations. In data preparation and batch analytics, they recommend Hadoop and sub-projects of Hadoop, such as Hive and HBase. In processing structured data Hadoop and Hive is an option, but they have found distributed analytical databases such as EMC Greenplum and HP Vertica useful for performance-related reasons and for integration – these can serve as backends for Business Intelligence (BI) software simplifying visual interaction. In processing semi-structured data their recommendations are various: HBase and Cassandra for hierarchical, key-value data organization, for graph analysis Neo4j and uRiKa and finally for geospatial data PostGis, GeoTools and ESRI. (Begoli & Horey 2012: 216–217).

Finally, the third principle of Begoli & Horey (2012), *Make Data Accessible*, unlike the previous two, concerns the results of KDD instead of the process itself. Based on their experience they deem paramount to expose the results with easy access and in an understandable form. Their three approaches to this are using open popular standards, light-weight architectures and exposing the results via API. (Begoli & Horey 2012: 216).

2.2 Hadoop Big Data platform

Apache Hadoop is one of the most used and well-known distributed computing platforms. It originated from a need to scale search indices at Yahoo! and was inspired by papers from Google describing their development of Map Reduce System and Google File System. First parts of Hadoop were created in aid of an open-source search engine software, named Apache Nutch, in 2005 and those parts became later an essential part of infrastructural software at Yahoo!. It soon became clear that this software could be much more than just a part of a search engine, that it was actually a generalizable distributed computation framework. Therefore, these components were separated into open-source project Hadoop. (Mendelevitch, Stella & Eadline 2017: 37–38; Mazunder 2016: 51).

From the first commits of the project in 2005, it took years for the platform to mature and evolve. Eventually, as the platform stabilized, several companies noticed the business opportunity around the framework: Cloudera was established in 2008, Hortonworks established by Yahoo! in 2011 and large IT-industry companies including EMC, Amazon, IBM, MapR, Oracle and Intel also entered the market (Mendelevitch et al. 2017:38; Ossous et al. 2017:14). Hadoop has evolved into a software ecosystem that can form the

essential parts of a data center operating system to scalably store, process and analyze Big Data. It consists of three main parts: a distributed file system, resource manager and distributed data processing frameworks. This is illustrated in Figure 4. Distributed data processing frameworks are located in the application layer. (Mendelevitch et al. 2017: 32–28).

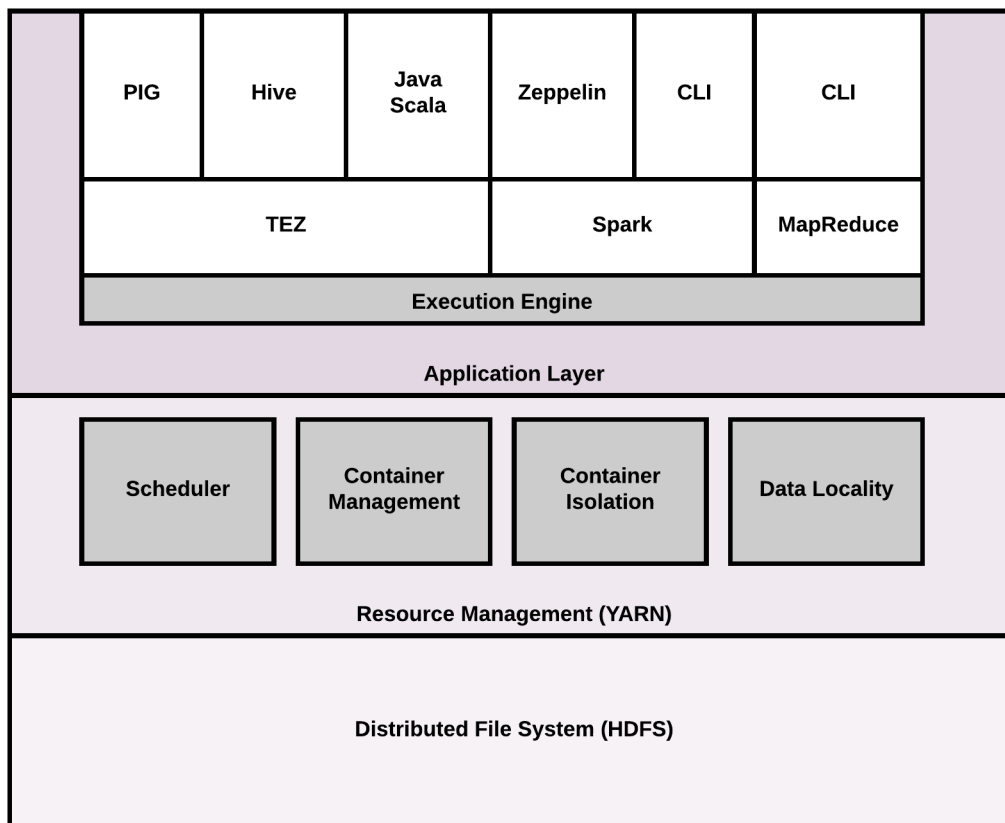


Figure 4. Hadoop architecture. (Adapted from Mendelevitch & al. 2017: 34; White 2015: 79).

As discussed in the section 2.1.2, Hadoop itself can be considered also as a core of BD platform architecture, with additional frameworks, modules, APIs and software connected to it. One well-known example of the larger implementation is the Berkeley Data Analytics Stack. (Mazunder 2016: 108).

2.2.1 Distributed file system

Hadoop is capable of operating with several different distributed file systems, for example, Amazon S3 and Microsoft Azure Blob storage system which are more suitable for cloud deployment, than the original Hadoop Distributed File System (HDFS) which is discussed here (White 2015: 53). HDFS is an open-source version of Google File System (GFS) developed by Google. HDFS is scalable, by built-in failure tolerance in the software layer, which in turn makes it possible to run it with less expensive (more fault-prone) commodity hardware resulting in cost-efficiency. HDFS can store large single files, even in terabyte sizes, and can store both unstructured and structured data. In HDFS the location of the data is communicated and the calculations are performed at the data. This helps to avoid unneeded network traffic, as only the calculations and results are transferred and it is congruent with the design goal of streaming data access: write-once, read-many-times. HDFS is aware of the network topology and always the fastest path to the copy of data is used. (Mendelevitch et al. 2017: 31–32; Oussous et al. 2017: 7; White 2015: 44, 70–71).

HDFS stores files in blocks, similar to single disk file systems. The default block size is 128 MB. If the file is smaller than the block size, only the needed amount of space is used. The large block size is due to the design goal of trying to minimize data seek times and the attempt to make the access time consist of as much as possible of the actual data reading and transferring. Therefore, reading a large file consisting of several blocks approaches the actual disk transfer rate. Many Hadoop installations use larger block sizes and as the transfer speeds of the disks grow, the default block size will be revised. (White 2015: 45).

Block abstraction has several benefits. Firstly, a single stored file can be larger than any of the hard disks used by the system, as it is stored as blocks on different nodes. Secondly, it simplifies both storage management where it is easier to calculate storage locations with fixed block sizes and file metadata issues, as the access to the file can be handled with another system as the blocks are just chunks of data. Thirdly, with blocks it is easier to cope with the replication of data. By default, the replication rate is three, meaning each file is cut into blocks and each of these blocks is stored three times to different disks and nodes. Therefore, with this setting, storing the file will take three times the size of the file in HDFS. (White 2015: 46).

In Figure 5 is presented the general HDFS workflow at an abstract level. NameNode is aware where each of the blocks is located in the system and which files they are parts of. If the client wants just a file list, it communicates only with the NameNode which provides the list. If a client wants to read or write data, NameNode tells the client the DataNode servers which contain the first few blocks in the file and thereafter the client communicates directly with these servers to access the data. These DataNodes are sorted by the NameNode by their proximity to the client. If a client itself is a DataNode and hosts a copy of the block, it will read from the local DataNode.

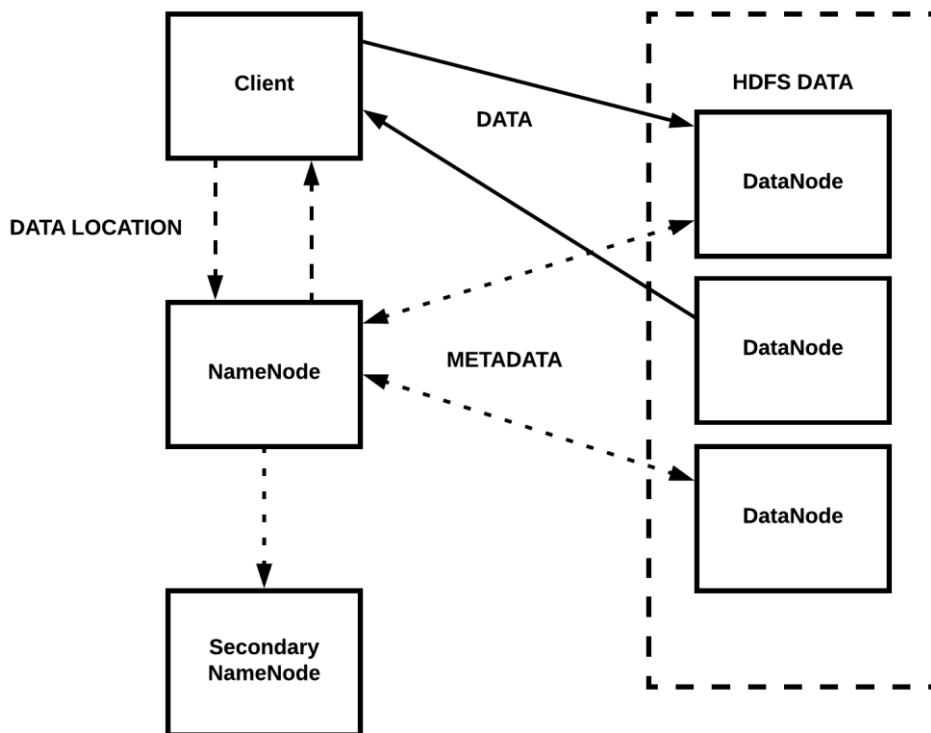


Figure 5. HDFS architecture operation. (Adapted from Mendelevitch et al. 2017: 33).

2.2.2 Resource manager layer

Apache YARN (Yet Another Resource Negotiator) was introduced in Hadoop version 2. In earlier versions of Hadoop MapReduce was both the application and the resource manager itself. It consisted of jobtracker and one or more tasktrackers. Jobtracker coordinated both scheduling and task processing while tasktrackers run tasks and communicated to jobtracker about progress. This older version is often referred to as MapReduce 1. Contrast to MapReduce 2 where the MapReduce is an application passing resource requests to YARN. YARN schedules tasks ensuring maximizing data locality and that system resources are utilized efficiently according to configured priorities. (Mendelevitch 2017: 34; White 2015: 79–84).

With the introduction of YARN, there were several design goals. Firstly, there was a need for multitenancy, to open up Hadoop to other distributed applications beyond MapReduce. This is achieved by an added layer of distributed execution engines such as Spark, MapReduce or Tez running as YARN applications on the resource manager layer. Applications such as Hive, Pig or Zeppelin interpret commands to the execution engines and do not use YARN API directly. YARN works by the execution engine contacting the ResourceManager (RM) to request it to run an Application Master (AM) process. RM finds a YARN NodeManager that is able to launch the AM in a container. AM can then depending on the application request more containers from RM or simply return the results back to the execution engine. AM schedules tasks, monitors TaskTrackers, maintains counters and restarts failed or slow tasks. Timeline server provides storage of the application history. AM lifetime can vary from one AM for one user job, one AM per user session of multiple jobs too long-running AM that is shared with different users. (White 2015: 80–85, Oussous et al. 2017: 8).

Secondly, there were performance-related reasons for re-design of the Hadoop architecture, more specifically scalability, availability, and utilization. YARN divides resources of the cluster, mainly CPU cores and memory, into containers that are isolated from the other users. As it is concerned with large volumes of data, YARN also controls data locality as a resource and can request a container run as close as possible to the source of the data. YARN also introduces user configurable schedulers to help with performance configuration. As in real-world clusters and use-cases are more or less unique, there are three schedulers available. FIFO (first in, first out) scheduler forms a queue of requests and runs them in order. Obviously, it is not well suited for clusters with multiple users or

user groups. With Capacity Scheduler system resources are divided by the user configured manner in queues, where a free queue is picked for new jobs. This leads to underutilization of the cluster resources as they are reserved for queues that are not necessarily in use. Fair Scheduler uses the system resources more dynamically, the principle is that all the system resources are allocated for the jobs running at the moment. If a new job starts, it is then allocated an equal share of the resources. Once a job finishes, resources it has used are then re-allocated to the still running jobs. This approach guarantees the full utilization of the system resources, the drawback being the delay and resources used for the re-allocations. (Mendelevitch et al. 2017: 34; White 2015: 84–87).

2.2.3 Application layer

Previous two layers discussed can be thought as a foundation, on which the actual application palette of Hadoop is built on a case by case basis. Organizational context, the requirements of the users, planned work processes, component compatibility, connections to outside system and motivations driving the design are the key factors deciding what exact components are chosen for the system. There exist many more Hadoop components and external integrations than is possible to go over in the scope of a thesis. Most relevant ones are introduced briefly. Some of the most important components are presented in more detail.

The first objective in BD Systems is the ingestion of Data. In order to process and analyze the data, it has to be collected into the system. Hadoop has several components for this task. Apache Sqoop is able to import and export data to any external data storage that has bulk data transfer capabilities with default and custom connectors, though usually it is used to bring in external data in Hive, for example (White 2015: 401–403). Flume is suitable for high volume transfers of external event-based data into cluster storage (White 2015: 381). Flume is a continuous stream processing system, but there exists a batch system based approach for the same problem space, Chukwa (Oussous et al. 2017: 9). Data can also be imported in the cluster manually in batches by copying it to HDFS.

Data needs to be stored inside the cluster for the cluster components to be able to process and refine it. Apache Hive is a data warehouse system running a top of Hadoop. The hive was originally developed at Facebook to allow data scientists to query massive amounts of collected data in familiar SQL by using HiveSQL. Hive functions both as a storage and analytics platform and can be connected to BI tools via ODBC connectivity. Hive is based

on the outside on familiar database schemas. Apache HBase is NoSQL column oriented key-value database designed for real-time read/write access on random datasets. (Oussous et al. 2017: 8, White 2015: 471; White 2015: 575).

Big Data can be seen with two different perspectives: data-in-motion and data-in-rest. Batch analytics is concerned with the former and streaming data solutions with the latter. For streaming data solutions there exist several components. Apache Kafka is a distributed streaming platform that is used to building real-time data pipelines between systems or applications, and is also used in support of other Hadoop components in batch data scenarios (Apache, 2018a). Apache Storm is designed to ingest data from various systems in real time, Twitter or Kafka for example, and write it to a variety of output systems (Mazunder 2016: 91). A relatively new project, Apache NiFi offers web-based UI for data routing, transformation and system mediation logic with directed graphs (Apache, 2018b). Apache Druid is an upcoming and developing component for storing, querying and analyzing large event streams currently in incubation phase (Apache, 2018c).

Data analytics and processing could be said to be the most important part of the platform and other components exist to make this possible. MapReduce was the original analytic tool in Hadoop. It is still powerful for parallel processing but as it requires programming skills and development time for developing and testing both custom map and reduce functions, it has shifted towards a language-in-the-middle, to which higher level languages are interpreted to (White 2015: 141). Apache Pig offers a layer of abstraction more compared to MapReduce, making the possible transformation of complex data structures with a language called PigLatin and offers web-based UI as a development platform while supporting external programs and not requiring a schema, supporting semi-structured and unstructured data (White 2015: 423; Oussous et al. 2017: 9). While Hive is also a data warehouse, it offers HiveSQL – a SQL like language which is parsed by MapReduce, Tez or Spark – for analyzation and transformation of data stored within, mostly in ELT (extract, load, and transform) use cases (Mazumder 2016: 58).

Apache Spark is best described as an open-source distributed Framework emphasizing speed by in-memory processing that was developed in AMPLab of UC Berkeley in 2009. Like other components, Spark has evolved since it was open sourced in 2010. The main abstraction Spark makes is called Resilient Distributed Dataset (RDD). RDD is a read-only collection of objects stored in system memory across multiple machines, on which

transformational logic can be applied in Scala or Python. On RDD a newer, more accessible abstraction has been built called DataFrame which makes usage more straightforward. There are four key features built around Spark. The first key feature is Spark SQL which unifies relational databases and RDD allowing users to perform queries both on imported datasets like Hive tables and data stored in RDDs or DataFrames. The second key feature Spark MLlib is a distributed machine learning framework built on top of Spark, offering for example regression models missing from Mahout. The third key feature is GraphX, which is a library for parallel graph computation built on top of Spark, extending the features of Spark RDD API. GraphX offers different operators that support graph manipulation and provides a library of common algorithms such as PageRank. The fourth and final key feature is Spark Streaming. Spark streaming is a component bit similar to Apache Storm, it provides automatic parallelization in addition to scalable and fault-tolerant streaming processing. Instead of normal Spark abstraction of RDD Spark Streaming uses a discretized stream called DStream. These discretized parts of the stream can then be processed. (Acharjya & Kauser 2016: 516; Oussous et al. 2017: 10; Mendelvitich et al. 2017: 42–43)

Even though Spark only offers Command Line Interface (CLI) Apache Zeppelin provides a web-based UI with deep Spark integration. Zeppelin is open-source multipurpose notebook supporting over 20 different language and software backends including Java, R, Python, Scala, SQL, Pig, SAP, and Mahout. It offers rapid data visualization, collaboration, sharing variables between Spark version of Python and R via ZeppelinContext. It is also usable in the Data Ingestion role. (Apache 2018d).

There exists other algorithm libraries in Hadoop ecosystem besides the mentioned GraphX and MLlib built on Spark. Apache Mahout is an open source machine learning software intended for creating models with machine learning algorithms, offering Java and Scala-based APIs to optimized algorithms developed by companies such as Google, IBM, Amazon, Yahoo, Twitter and Facebook (Oussous et al. 2017:11, Mazunder 2016: 61). Apache DataFu provides two libraries: Apache DataFu Pig and Apache DataFu Hourglass (Apache 2018e). DataFu Pig is a collection of tested user-defined functions to Pig and DataFu Hourglass incremental processing framework for sliding window calculations (Apache 2018e).

To ensure proper working of a large cluster, supporting components are needed. Apache Zookeeper ensures reliable distributed coordination of applications and clusters, by

providing centralized in-memory services for example configuration information and naming and is used in providing high availability for ResourceManager (Oussous et al. 2017: 12; Mazunder 2016: 65). Apache Oozie is open-source workflow scheduler system designed to manage various types of jobs needed to implement a data processing pipeline, working by creating a Directed Acyclical Graph (DAG) out of workflow jobs (Oussous et al. 2017: 12; Mazunder 2016: 66).

Access control and security are essential in a multiuser environment. Hadoop is still not completely matured in regarding security. User security consists of both authentication and authorization. User authentication can be done via Lightweight Directory Access Protocol (LDAP) connecting with System Security Services Daemon (SSSD) connecting the Linux OS with LDAP. Hadoop supports Kerberos authentication for communication between the nodes of the cluster. Apache Knox is REST API based gateway providing a single REST access point, while also complementing Kerberos secured cluster. Apache Ranger offers complete authorization service for Kerberos secured cluster. Access control can be fine-tuned on the very granular level on multiple services or actions, based on roles or attributes, including HDFS file access, access roles on different Hadoop components and auditing provided via Apache Solr. Unfortunately, Ranger has still deficiencies in components it supports. (Mazunder 2016: 62–63).

3 VALUE SENSITIVE DESIGN

Hoven (2013: 78) sees Value Sensitive Design (VSD) as the culmination of a development that started at Stanford in 1970s. There the moral issues and values embedded in technology were a central aspect of study in Computer Science and since then there have been several encapsulations of the principles. Hoven recognizes VSD as formulated by Friedman, Kahn & Borning (2008), as one of the first frameworks concerned on integrating values to design process and sees that other frameworks have later emerged, such as Values in Design and Values for Design. Manders-Huits (2011: 273) describes VSD emerging from studies regarding Human-Computer Interaction (HCI), which is congruent with the view of the evolution of VSD that Friedman, Kahn & Borning (2002: 1) present.

Friedman et al. (2008: 71; 2008: 85) see Computer Ethics, Social Informatics, Computer-Supported Cooperative Work, and Participatory Design as related approaches to VSD. In this thesis VSD was chosen as kernel theory directing the study as it has widespread usage in different fields of ICT for example Johri & Nair (2011), Mok & Hyysalo (2018), Dadgar & Joshi (2015), Xu, Crossler & Bélanger (2012), Wynsberghe (2013), Alshammari & Jung (2017), and Miller, Friedman, Jancke & Gill (2007). As the framework has evolved during a longer period, there exists constructive critique such as Manders-Huits (2011), Jacobs & Huldtgren (2018) or Borning & Muller (2012), which provide additional guidance on implementation.

Friedman et al. (2008: 69) define VSD as “theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process”. They see it as a tripartite methodology consisting of conceptual, empirical and technological investigations. These are discussed further on following sub-chapters. All three are iterative processes, affecting each other during the course of the research. Essential to the practice of VSD is identifying stakeholders, defined as users of the system and indirect-stakeholders, defined as people affected by the new system, researching what kind of values all of them hold and how the actual technological design can then take into consideration these values (Friedman et al. 2008).

There are eight central unique features in VSD according to Friedman et al. (2008: 85–86). Firstly, VSD attempts to influence the design of technology early in and throughout the design process. Secondly, VSD is implementable in other arenas besides the work-

place. Thirdly, VSD contributes a unique tripartite methodology which is applied iteratively and integratively. Fourthly, VSD incorporates all values, especially those with moral import. Fifthly, VSD distinguishes between usability and human values with ethical import. Sixthly, VSD identifies and analyses two sets of stakeholders, direct and indirect. Seventhly, VSD is integrational theory and values are not viewed either as inscribed into technology nor simply as transmitted by social forces. Eighthly, VSD is grounded on the proposition that “certain values are universally held, although how such values play out in a particular culture at a particular point in time can vary considerably” (Friedman et al. 2008:86). (Friedman et al. 2008: 85–86).

In the center of the VSD process are the values. Friedman et al. (2008: 70–71) explain their definition of value being a broader term, referring to what person or group consider important in value, which is based on the Oxford English Dictionary definition. They acknowledge the problematics and variation of the relation of values and ethics, and ultimately they depend on the distinction between fact and value, where facts do not logically entail value. “Is does not imply ought” Friedman et al. (2008: 71) which is known as the naturalistic fallacy. Further, Friedman et al. (2008: 71) continue “values cannot be motivated only by an empirical account of the external world, but depend substantively on the interests and desires of human beings within a cultural milieu”. Values in the context of VSD can be described as “what a person or group of people consider important in life” (Friedman, Kahn, Borning & Huldtgren 2013).

3.1 Investigation types in Value Sensitive Design

Friedman et al. (2008: 71–72) describe the application of the three types of investigations in different research projects comparable to paintings. In paintings created by various authors, different techniques are applied with a multitude of ways to form a whole, which is more than the sum of the parts, and still dissimilar to another painting. “The diverse techniques are employed on top of the other, repeatedly, and in response to what has been laid down earlier” as Friedman et al. (2008: 71–72) describe it. Next, these investigations are discussed further.

3.1.1 Conceptual Investigations

Conceptual investigations in VSD consist of finding out the direct and indirect stakeholders, how they relate to the system and how they are affected by it, what kind of values are implicated and how the design decisions and trade-offs between competing values should be handled. Additionally, Friedman et al. (2008) see that by conceptualizing of specific values carefully can fundamental issues related to the project be found and identified, which in turn can provide a basis for comparing results between different research teams. (Friedman et al. 2008: 72)

Friedman et al. (2008: 87) define direct stakeholders as those, “who interact directly with the technology or technology’s output” and indirect as those, “who are also impacted by the system, though they never interact directly with it”. Further, Friedman et al. (2008: 87–88) point out that it within both groups of stakeholders, several subgroups may exist and one individual may be part of more than one stakeholder group of a subgroup. According to Friedman et al. (2008: 88), organizational power structure does not follow the division to direct or indirect stakeholders, so the effect of it needs to be carefully considered.

After identifying stakeholders, Friedman et al. (2008: 88) suggest to identify benefits and harms for each stakeholder group. Friedman et al. (2008: 88) present three suggestions to attend to. Firstly, benefits and harms will vary for each indirect stakeholder and more complex system, the larger group of people it affects – consider the World Wide Web for example. Friedman et al. (2008: 88) suggest in such situations to give priority to indirect stakeholders who are strongly affected or to large groups that are somewhat affected. Secondly, Friedman et al. (2008: 88) see the necessity to attend to the issues of technical, cognitive and physical competency. Interests of such groups should be attended during the design process by representatives or advocates. Thirdly, they suggest personas as an investigation tool for the benefits and harms of each stakeholder group. Friedman et al. (2008: 88) point out that with using personas one has to be careful not to reduce them into stereotypes and that in VSD one persona can be a member of several stakeholder groups. (Friedman et al. 2008: 88).

Once benefits and harms to each stakeholder group are identified, these should be mapped to corresponding values (Friedman et al. 2008: 88–89). Friedman et al. (2008) note that mapping can be relatively straightforward but it can also be less direct and multifaceted.

They cite an example of mood improvement, which benefit can potentially implicate creativity, productivity and physical welfare in addition to the psychological welfare.

Once the key values are identified, a conceptual investigation of each should be performed with the aid of relevant literature according to Friedman et al. (2008). Friedman et al. (2008: 89) point out that “philosophical ontological literature can help provide criteria for what value *is* (cursive added)” and help in the following empirical investigations. Potential value conflicts should be examined next after the values have been defined. Friedman et al. (2008: 89) see value conflicts more as restrictions on the design space instead of choosing one over another, even though often supporting value might hinder support of another. (Friedman et al. 2008: 89).

3.1.2 Empirical investigations

Friedman et al. (2008: 72) recognize the limitations of conceptual investigations and see that information provided by empirical investigations targeting the human context of the technological artifact is critical for many analyses. Secondly, they see the value of the empirical investigations regarding the evaluation of the success of the design. With empirical investigations, Friedman et al. (2008: 73) suggest researching for example how stakeholders apprehend individual values in the interactive context, how they prioritize design trade-offs between competitive values and prioritization of individual values and usability considerations. Additionally, Friedman et al. (2008: 73) recognize that technological artifact has an effect on organizations as well as single stakeholders, organizational value considerations affecting the design process can also be examined, for example, organizations’ motivations, methods of training and dissemination, reward structures, and economic incentives. (Friedman et al. 2008: 72–73).

According to Friedman et al. (2008: 72), empirical investigations can be applied to any human activity that can be observed, measured or documented. Therefore, they suggest that suitable methods consist of all quantitative and qualitative methods available in social science research, including for example observations, interviews, surveys, experimental manipulations, collection of relevant documents and measurement of user behavior and human psychology.

3.1.3 Technological investigations

VSD proposes that technology itself, by the properties of technology, provides different value suitabilities and thus supports some values and activities based on those values more readily than others (Friedman et al. 2008: 73). First of the two possible forms of technological investigation in VSD examines this. In this form of technological investigation is researched how the existing technology supports or hinders certain values, what kind of design trade-offs between values exists. The second form of technological investigation, implemented in this thesis, takes a different approach and it is considered with design and development of technology from a value perspective – how to design technology to support the values identified in the conceptual investigation. Friedman et al. (2008: 73) note that technological investigations may appear similar to empirical investigations, but they differ by their unit of analysis. The technological investigation is concerned with the technology while empirical investigations are concerned with the social units that are affected by technology. (Friedman et al. 2008: 73).

3.2 Critique of Value Sensitive Design

One criticism of VSD is directed at the definition presented. Mander-Huits (2011: 279–280) claims that exactly what happens with VSD framework, a conflation of facts and values if values are “taken as the normative input for the VSD of a technology”. Further, Mander-Huits (2011: 280) sees that Friedman et al. (2008) claim of values “depending substantively on the interests and desires of human beings within cultural milieu” actually implies a sociological conception of values rather than ethical one. Jacobs & Huldtgren (2018) proposes that to avoid this naturalistic fallacy an ethical theory should guide the examination of values to form the normative input for the VSD analysis. They see that ethical theory could provide arguments for value prioritization and for the trade-offs that eventually raise in system design and propose that especially mid-level ethical theories are well suited in the VSD process. Mander-Huits (2011: 282–283) also points out the need for an ethical theory for exactly these mentioned considerations.

Friedman et al. (2008) proposal that some values are universally held and can thus provide normative direction is recognized also by Borning & Muller (2012) as problematic, and

their suggestion is in the instances of applying VSD to use qualifying prescriptive statements (if we want to support x, then we should do y) and in the instances of using VSD to be clear about the position of the researchers and their commitments (Borning & Muller 2012: 1125–1127). Related is their suggestion to provide cultural and viewpoint context for the typical list of values presented in VSD papers (Borning & Muller 2012: 1125). Additionally, they do not see enough voice of the participants in the publications of VSD. Borning & Muller (2012) claim that there is overclaim of authority and knowledge when substituting voice of the researcher over the participants and see that there are value and fidelity in allowing the participants to speak for themselves (Borning & Muller 2012: 1125). Lastly, they recommend making more salient the voice of the researcher when writing about VSD research thus allowing the reader to see more clearly researchers own culture and assumptions (Borning & Muller 2012: 1126).

Partly additional criticism by Mander-Huits (2011) is concerned with identifying stakeholders. Especially she sees recognizing parties affected by more complex technologies increasingly difficult, especially so when considering in-direct stakeholders (Mander-Huits 2011: 277–278). Yoo (2018) suggest a nascent method for this. Lastly, Mander-Huits (2011) points out that empirical methods employed in VSD to gather the stakeholder values need a lot of consideration (Mander-Huits 2011: 278–279). Essentially her critique is based to the difficulty of stakeholders correctly assessing the new technologies, the values stakeholders communicate could be experienced and interpreted differently what they intended (thus resulting in different norms and actions than actually implied) and the fact that values change. Mander-Huits (2011: 279) recommends that the method used for value elicitation would be deliberative than taking a form of a survey.

4 DESIGN SCIENCE AND RESEARCH PROCESS

“Scientists try to identify the components of existing structures, designers try to shape the components of new structures” (Alexander, 1964)

Design science can be described as a search path of Simon (1996) exploring methodologies to combine the two roles Alexander mentions in the quote above. To bring scientific rigor into design, to allow evaluations and comparisons of design solutions and to generate the necessary ontology for such discussion. Design science is discussed in detail in this chapter and the combinatory approach of application of both VSD and DS in the thesis is detailed.

4.1 Design science

Cross (1993) sees the roots of design science in the series of conferences held in the 1960s and 1970s related to *Design Methods Movement*. *The science of Design* influenced especially by Simon (1996) with his *The Sciences of the Artificial* first published in 1969, Cross (1993: 21) sees as “body of work which attempts to improve our understanding of design through ‘scientific’ (i.e., systematic, reliable) methods of investigation”. *Design science* (DS), on the other hand, Cross (1993: 21) defines as “explicitly organized, rational and wholly systematic approach to design; not just the utilization of scientific knowledge of artifacts, but design in some sense as a scientific activity itself”.

Iivari (2008: 40) views Information Systems (IS) as an applied science, as do Peffers et al. (2008: 46) and elaborate that often in IS theories from other disciplines are applied to solve problems in the context of information technology and organizations. Iivari (2007: 40) claims that much of the early research performed in IS was DS, that it was focused on system development approaches and methods, for example, socio-technical approach and the infological approach actually being Design Science Research (DSR). Mandiwilla (2015: 315–316) sees the current state of DSR as positive, that Peffers et al. (2008) have outlined the overall process and others have shown that DSR can operate in field settings addressing relevant problems and DSR knowledge contribution has been further expanded in the areas of predictive and explanatory theories. Mandiwilla (2015: 316) sees the work of Gregor and Hevner (2013) important as outlining DSR knowledge contribution based on the level of abstraction and maturity. Further Mandiwilla (2015: 316) see

importance in the work of Gregor and Hevner (2013) as they discuss how the relationships between descriptive and prescriptive knowledge can influence different forms of DSR.

Gregor and Hevner (2013: 343) divide useful knowledge in DSR into two types, descriptive knowledge (Ω) and prescriptive knowledge (Λ). Descriptive knowledge encompasses natural, artificial and human phenomena and what we know of them – what kind of laws govern them and the regularities among the phenomena (Gregor & Hevner 2013: 343). Gregor & Hevner (2013: 343) describe it as “what” knowledge. Iivari (2007: 46) defines descriptive knowledge as aiming to describe, understand and explain how things are. Prescriptive knowledge, on the other hand, consists of constructs, models, methods, instantiations and design theories (Gregor & Hevner 2013: 343). It is the “how” of human artifacts, as Gregor & Hevner (2013: 343) describe it. Iivari (2007: 46) sees prescriptive knowledge as interested “in how things could be and how to achieve the specified ends in an efficient manner”.

The essential idea of how DSR process should proceed according to Gregor & Hevner (2013) is presented in figure 5. They see the starting point as an important opportunity, a challenging problem or insightful vision or conjecture for something innovative which is then transformed into research questions. Gregor & Hevner (2013: 343) see that when first considering the research questions, first questions raised would be “What do we know of this already?” and “From what existing knowledge can we draw?”. Next would be the investigation and exploration of both knowledge bases Ω and Λ . From the Ω base, relevant knowledge may be found in different elements such as justificatory theory relating to the goals of the research. From the Λ base investigation should be directed at known artifacts and design theories that have been already used to solve similar research problems. Gregor & Hevner (2013: 343) state the objective here being to provide a baseline of knowledge on which to evaluate the novelty of the new artifacts and knowledge resulting from research. They see it to be common for the new design research contribution being either an important extension of an existing artifact or an application of the existing artifact in a new application domain. (Gregor & Hevner 2013: 343).

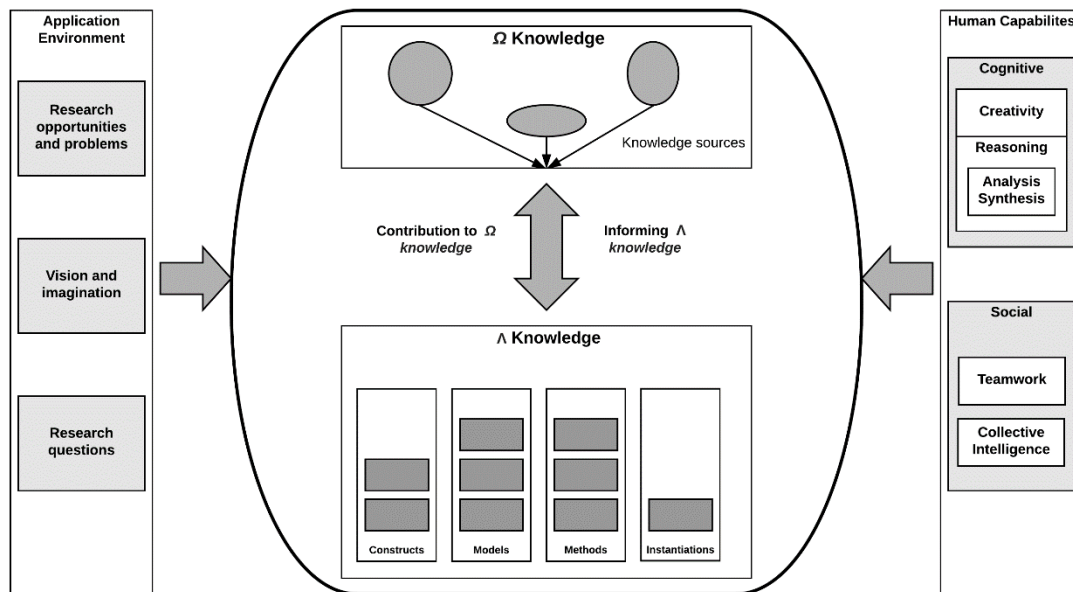


Figure 6. Roles of knowledge in DSR. (Gregor & Hevner 2013: 344).

Earlier, in 2007 Hevner described DSR process through three connected iterative cycles, based on and in agreement with the twelve theses of Iivari (2007: 56). The first cycle is the relevance cycle, which grounds the research to a relevant real-world context. It defines the problem and the validity of the solution. The second cycle, the rigor cycle is where the existing knowledge is examined to ensure the rigor of the design, by ensuring the novelty of the design. Hevner (2007: 4) sees the requirement of the kernel theory or the grounding theory based on descriptive knowledge as unnecessary in every situation, and the requirement even hindering the goals of DSR. Moreover, and importantly concerning research conducted in this thesis, he would include creative insights as a source for grounding the research. The third cycle, the design cycle is based on the Generator-Test Cycle of Simon (1996: 129). Within design, alternatives are created and evaluated until a satisfactory solution is reached. Hevner (2007: 4) describes input to this cycle as the context from relevance cycle and the evaluation methods from the rigor cycle. According to Hevner (2007: 5), it is important to maintain the balance of these design and evaluation activities.

4.2 Research process

DS is applied in this thesis according to the Design Science Research Methodology (DSRM) proposed by Peffers, Tuunanen, Rothenberger, and Chatterjee in 2008. Peffers et al. (2008: 47) saw the need for a common framework for DS research in Information Systems as it seemed *research of design* was a minority amongst published papers and mainly was used and accepted in other disciplines, such as engineering. In engineering Peffers et al. (2008: 47) noted research of design was dominated by research streams such as requirements engineering and software engineering. Thus, Peffers et al. (2008) propose that a common framework is required for the readers and reviewers to recognize and evaluate the results of such DS research in IS. As such methodology to perform DS research and to present it did not exist in literature, they proposed DSRM.

DSRM consists of six activities. The first activity is problem identification motivation, where the specific research problem is defined and value of the solution justified. Second activity consists of defining the objectives of the solution, where the objectives of the solution are inferred from the problem definition and knowledge of what is possible and feasible. The third activity is called design and development, consisting of creating the artifact. The fourth activity consists of a demonstration, using the artifact to solve a problem. Fifth activity is an evaluation, consisting of observing and measuring how well the artifact supports a solution to a problem. Sixth and last activity is communication, consisting of communicating the problem and its importance, the artifact and its utility and novelty, the rigor of its design and its effectiveness. Activities are not necessarily performed in the nominal order and the whole process can be iterative, the entry point to the process can also be any step between first and fourth, those included. (Peffers et al. 2008: 52–56).

Within activities of DSRM, the three concepts of VSD are performed iteratively, technological, empirical and conceptual investigations as pictured in Figure 7. As case project can move in unanticipated directions, technological assumptions can prove incorrect, new issues and limitations can emerge, the iterative nature of VSD is well suited for the study: *“The diverse techniques are applied on one top of the other, repeatedly, and in response to what has been laid down earlier”* (Friedman et al. 2008: 71–72).

DSRM Phase		Meta-Action(s)	VSD Phase	Actions in this study
Identify the Problem	Literature review	Start with Value, Technology or Context of Use	Research related to Big Data Research related to Hadoop Research related to SESP-project Research related to formation of knowledge from data Research related to Organizational context of the project Research related to DSR and VSD methodology	
	Empirical Investigation (First)	Setting of an objective	Specification for BD platform provider Case onboarding and reading of case related documentation Identification of design principles as the goal of research Planning of research	
	Technological Investigation (First)	Building of first prototype	Demonstration of the first prototype via Reddit analysis	
	Empirical Investigation (Second)	Identify Stakeholders	Stakeholder identification by Stakeholder token method Securing of system resources Design of a prototype	
	Technological Investigation (Second)	Conduct a conceptual Investigation of Key Values Identification of Potential Value conflicts	Identification of initial key values identified in first empirical investigation Research related to empirical data gathering Design of the interviews	
	Empirical Investigation (Third)	Map Benefits and Harms onto Corresponding Values	Performing the interviews Analysis of the data Benefits and harms identification	
	Empirical Investigation (Fourth)	Conduct a conceptual Investigation of Key Values Identification of Potential Value conflicts	Workshop related to platform organizational context Mapping of benefits and harms to values Conceptual investigation of key values by interview results Value conflict identification	
	Creation of artefact	Integrate Value Considerations into One's organizational structure	Creation of design principles	
	Creation of instantiation of the artefact		Implementation and iteration of the Design Initial performance testing Hardware upgrades Continuous development according to design principles Continuous design and evaluation of the platform	
	Evaluation	Positioned to future research		Documentation of the process
Thesis publication			Documentation of the future plans and case related information	

Figure 7. Research process used in this study.

5 DESIGN AND DEVELOPMENT

Challenges of the design of things are universal according to Simon (1996), if they are thought of as an artifact interfacing outer and inner environments. In this case, both of these environments are extremely complicated and full of possible design paths. As this case study was conducted in a relatively limited time period, it required of an alternative view of a goal – instead of the impossible task of exhausting all possible design paths, a guide map for the future was created. This guide map, the design principles, would serve the evolving platform in the future as each new crossroads of technological, financial or organizational opportunities or challenges was encountered. This alternate setting of a goal, seeing the design problem differently and changing the problem representation, is also something that Simon (1996: 131–134) proposes.

In this chapter is depicted this search for that alternate goal, consisting of the iterative VSD investigations and thus constructing the third activity of DSRM method, the design, and development and it is presented in chronological order.

First empirical investigation, consisting mostly on onboarding to the project, contact and communication with various actors in the project, reviewing of the existing project documents and writing of the initial specification for the original platform provider, is here left uncovered as it is part of the first DSRM phase, the identifying of the problem, and for reasons of clarity and brevity. It cannot be described as a rigorous or well-structured process. The knowledge gained, however, was used in the initial conceptual investigation of values and value-conflicts identified in the first empirical investigation.

5.1 Smart Energy System Research Platform -project

The research in this thesis is conducted as a part of the Smart Energy System Research Platform project (SESP). SESP is a highly ambitious two-year project that is motivated by the developments and changes in the fields of energy production and distribution, mainly focused around the concept of Smart Grid (Antila, Virrankoski, Kauhaniemi, Vartiainen, Larimo, Rajala, Galkina, Kock, Björk & Norrgrann 2016). Smart power grids are defined by Moreno-Garcia, Moreno-Munoz, Pallares-Lopez, Gonzalez-Redondo, Pala-

cios-Garcia, and Moreno-Moreno (2017: 45) as power grids that integrate various technologies, for example, electric vehicles and smart home appliances, and integrating methodologies such as demand response programs and decentralized power management of renewable resources.

There are four main objectives of SESP-project defined by Antila et al. (2016: 4), each started during the project and to remain active after the project. Firstly, a new electric systems laboratory, secondly actively updating Big Data Warehouse related to Smart Grid which is main theme of research of the thesis, thirdly new living lab sites that are part of real energy systems but connected to aforementioned energy laboratory and fourthly, new business concepts and models that are based on the utilization of energy systems Big Data, including both quantitative and qualitative data. (Antila et al. 2016: 4).

Commercial aspects of the SESP-project relate to studying of changes in the energy choices and behavior of both customers and citizens by gathering, combining and analyzing data provided by different in-depth studies concerning household behavior, renewable energy production plants and different R&D platforms. By refining and analyzing this data it is intended to design and develop customer driven market offerings, in order to facilitate real energy transition to renewable energy sources. Big Data platform designed and documented in this thesis is central to the process. (Antila et al. 2016: 5).

SESP-project provides several pragmatic limitations to the possibilities of design and development in the project. One of the risks evaluated in the research plan (Antila et al. 2016: 7), the reliability of the external service provider was realized quite late in the project process. As a result of late exit, recovery actions were severely limited by the time available, financial resources as the external partner would have had the responsibility of providing most of the infrastructure, and lastly, there were limitations of staff competencies as external domain experts were not available. Domain expertise had to be researched and, in addition, discovered by empirical tests of actually building Hadoop cluster prototypes. There were also changes in key personnel in the project that could have had some effects on the project process. The research in the thesis and building of data platform is conducted in the framework of SESP-project, goals of the SESP-project and restrictions the project provided.

5.2 Phase I: First technological investigation

The first technological investigation was required as Big Data analyses, in general, are constantly developing and especially technological solutions are ever evolving, a representational example being the multitude of the components existing and raising in the Apache Hadoop ecosystem. As of writing this, the number of different components amounts to 149 different projects (Roman 2018). Constant evolution and competition lead to a situation where best practices must be custom tailored for each situation. Understanding of specifics of the situation and understanding of the technology must be gained empirically by first hand. If one does not understand the technology package, how it interacts internally and how it can be interacted with from outside, it is not possible to understand what values are involved in practice.

First technological investigation constituted of building a working Hadoop cluster and examining the practicalities involved in the building of a Big Data platform. Additionally, a small demonstration was performed with analysis of semi-structured data to gain insights into how the system can be operated and what are the limitations.

5.2.1 The building of the first prototype

Initial testbed exploring the Hortonworks Data Platform (HDP) distribution of Hadoop, was built earlier during the SESP-project before this researcher was involved in the project. It was built on a collection of older computers with old hardware with severely limited capabilities, serving mainly as a test bed of the feasibility of utilizing HDP. As a result, HDP was chosen as the backbone the platform was to build around of. At this stage, an outside platform provider was also involved and recommending HDP.

First actual and fully working was built during the research process on virtual machines running on Microsoft Hyper-V version 10.0.14393.0. These virtual machines ran on a single machine, with one Intel Xeon Silver 4114 CPU with 10 cores and supporting HyperThreading, resulting in 20 countable cores. The system had a total of 96 GB of memory and 3 TB of usable hard disk space. As there were other users of the system, for the prototype cluster usage was allocated a total of 80 GB of memory and a total of 2.9TB disk space. One investigation venue was how to maintain and manage several nodes, therefore a total of 10 machines was planned to be included in the cluster. Detailed spec-

ifications can be found in table 2. This division of resources was done based on the resources available and on the knowledge of practitioners available in various forms consisting of comments, experiences and documented best practices. Final specification and configuration were reached after a few iterations.

Table 2. Node specification of Prototype 1.

Machine name	Role	CPU/HDD/Memory	Components
cluster2ms1	Main Server	4 Cores/250GB/12GB	26
cluster2ms2	Secondary Server	4 Cores/250GB/12GB	External – Puppet, Kerberos, LDAP and utilities
cluster2sl1	Worker	2 Cores/300GB/4GB	13
cluster2sl2	Worker	2 Cores/300GB/4GB	19
cluster2sl3	Worker	2 Cores/300GB/4GB	5
cluster2sl4	Worker	2 Cores/300GB/4GB	24
cluster2sl5	Worker	2 Cores/300GB/4GB	6
cluster2sl6	Worker	2 Cores/300GB/4GB	6
cluster2sl7	Worker	2 Cores/300GB/4GB	6
cluster2sl8	Worker	2 Cores/300GB/4GB	24

Main benefits of building, designing and testing of prototype were better understanding of the practicalities involved in configuring different components of Hadoop ecosystem, how the system performed under load, what kind of resources are actually needed for full scale system, understanding of the maturity level of the technology in general and especially of some of the components. Actual testing with importing, processing and analyzation of datasets were considered one of the main contributions of building prototype 1, as it allowed the designers of the system to gain a better understanding of the availability of different languages, APIs, IDEs and capabilities related to planned processes. Many details of implementation were discovered during prototyping that should be designed and configured differently for the next prototype. Discovery of the need for more computational resources consisting of more CPU cores, additional memory, and storage space could have been described as one of the most significant results.

Moreover, utilization and application of Kerberos authentication and Ranger access control were tested, with tying the user access to a created LDAP server user roles and accesses. To investigate the problem area of managing multiple clusters automatically, the secondary server was dedicated as a Puppet and utility server. Foreman web-based UI was installed to improve the usability of the said server and custom made Puppet scripts were created to handle some limited parts of the deployment process and to test automation in practice.

Discovery of the technological limitations of the components and how they could interplay, how compatible they were in the real-world situation provided extremely valuable insight in how different values could be either be supported or diminished by design decisions in practice.

5.2.2 Prototype demonstration

As a proper test of the capabilities of the system and the limitations in more complex usage, a text analysis test was performed. A dataset consisting of a collection of Reddit messages that was publicly available (Pushshift 2019) was utilized in the test. Because of the limitations of hard drive space on the prototype, dataset utilized in the demonstration was limited to all public messages written in Reddit during the year 2017, consisting of approximately 970 million messages. As a demonstrative research question, messages mentioning the word “energy” were searched for words mentioning renewable energy, and different renewable energy forms were then ranked based on the counts they were mentioned.

The transformations of the messages were done with Scala utilizing Spark on Zeppelin notepad. Messages were lowercased, tokenized and sanitized iteratively. HiveSQL was used in ranking and ordering of the results. These results and the process of how they were achieved was demonstrated for the board of the SESP-project.

By performing the demonstration, it was discovered that the platform prototype was capable of some of the planned uses and the available variety of analytical methods supported this kind of research. Experience-based knowledge of the ways to import data to the system, additional required configurations changes, and what the limitations of the system meant in practice was gained. Technical knowledge and experience gained by

developing and building the first prototype, and additionally by performing the demonstration would help in considering how different values could interplay in actual operations.

5.3 Phase II: Second empirical investigation

Friedman (2008: 87) suggest that the identification of direct and indirect stakeholders should happen during the first initial conceptual investigation. However, the critique by Manders-Huits (2007) and resulting proposed method to address this by Yoo (2018) led to a different approach in this study. Identification of stakeholders was performed empirically instead of just by considerations by a researcher in a conceptual investigation.

Manders-Huits (2007: 277–278) states that identification in any stakeholder analysis is an issue with critical importance. She sees the issue not only related to VSD, but if the aim is to employ VSD to design technological systems that are sensitive to human and moral values and a representative result of stakeholders is not reached, then everything building on it by further investigation into stakeholder values might be questioned. Mander-Huits (2007: 227–278) especially points out how this problem increases in complexity, especially related to finding out indirect stakeholders, as the complexity of technology in question increases.

5.3.1 Stakeholder tokens method

Yoo (2018) sees most VSD studies providing a list of stakeholders, while the detailed explanation of the method or methods used in identifying and selecting stakeholders are missing. She proposes a nascent method called Stakeholder Tokens (ST), which is a “concrete method and tool for conducting stakeholder analyses both with designers and with potential stakeholders” (Yoo 2018: 2). Yoo (2018) describes the ST method having three main goals, firstly a generation of a more inclusive set of stakeholders by uncovering easily overlooked stakeholders. Secondly, providing a more robust set of stakeholders by providing a rationale for their inclusion and thirdly, to clarify stakeholder dynamics within a complex socio-political setting. (Yoo 2018).

Yoo (2018) claims also ST having three distinct experiential characteristics. Firstly, the method is participatory and embodying principles of participatory design. Secondly, she sees the method being visual and tactile, “a tangible tool that helps participants to make sense of their mental models of complex networks and relationships” (Yoo 2018: 2). Thirdly she states the method being creative and playful and thus employing principles of ludic design. (Yoo 2018: 2).

ST proposed by Yoo (2018) has five steps. The first step is selecting the participants and Yoo (2018) suggest that researchers that are familiar with the topic of the research, create an initial list of both groups of stakeholders, direct and indirect. Based on this list participants to ST method should be chosen from the subset of stakeholder groups, considering resource constraints. The second step consists of choosing proper tokens for representing the stakeholders; a good token fits to one’s hand, is easy to move around the tabletop, form of the token is intuitive and familiar and avoids representing the stakeholder in a stereotypical way. The third step is the creation of list and labels, which participants may do either individually or divided into groups. Yoo (2018) suggest using prompting questions to get the participants to think of those who are central to the issue and to those who might be more difficult to uncover. The fourth step consists of attaching the labels to tokens. In the fifth and last step labeled tokens are placed on a sheet of paper and interrelationships among those stakeholders is drawn. Yoo (2018) also notes of using tokens to act out some of these relationships. (Yoo 2018).

5.3.2 Stakeholder identification

The method proposed by Yoo (2018) was employed in the research as a tool to discover stakeholders, both direct and indirect, and to gain an understanding of the relationships involved. As a preliminary list of stakeholders was evaluated with regards to resource limitations, a rather limited approach was chosen. Participants included author himself, both in a participatory roles as well as a facilitator, and another designer of the system, who was previously identified by the author as a member of several potential stakeholder groups, both direct and indirect. He had an understanding of both the organizational environment of the system, technological solutions and capabilities, and the context of the case project. Prototype nature of the system and the understanding of the technology and the resulting implications as discussed by Manders-Huits (2007: 278) was thought to limit understanding of the system as a whole of several potential participant groups, smaller participatory attendance was deemed sufficient.

The method was applied in a slightly modified form from the proposal of Yoo (2018). As a prompter, a high abstraction of the system and the external system environment was drawn and discussed. Some prompting questions such as “Who this system would affect?” were asked. Both participants then wrote down all stakeholders, without labeling stakeholders direct or indirect, to sticky notes individually. Internal University of Vaasa stakeholders were written down to blue notes and stakeholders considered external in some sense, were written down to red notes. Next in alternating order one of the participants read out a stakeholder, attached it to another piece of paper and the other participant then attached his related stakeholders next to it, removing the doubling of identical stakeholders. During this process stakeholders were discussed briefly. Next step was writing down the stakeholders to adhesive tape and attaching them to chess pieces used as tokens. It is possibly noteworthy that participants chose larger, more valuable pieces to represent both the Ministry of Education and Culture and the leadership of the university.

Once this process was done, two large pieces of paper were combined and an abstraction of the system was drawn in the middle. In co-operation, asking questions and discussing, then the prepared tokens were placed and relationships between them were described by writing next to the arrows drawn to represent the relationships. Some tokens were moved a few times and if several tokens were found to have almost identical relationships or context, a box representation of the context was drawn. Once every token was placed, a brief discussion took place where it was pondered if every key stakeholder is present and if the picture matches reality. Once there was not anything to add, photographs were taken and tokens removed. Once token was removed, stakeholder it represented was written to the place of the token and the resulting diagram stored for analyzation. The process of application of the modified method took approximately 1.5 hours.

The end result of the process is featured in figure 8 as a cleaned drawing, with omitted textual representation of the relationships for clarity purposes. Blue background denotes internal actor to University of Vaasa (UVA) and green denotes actor mainly considered outsider to University. Appendix 1 contains the original picture figure 8 is based on. Actors outlined with red were chosen for an in-depth interview. Criteria for selection was realistic accessibility, number of connections to other stakeholders and variety in perspectives related to the platform.

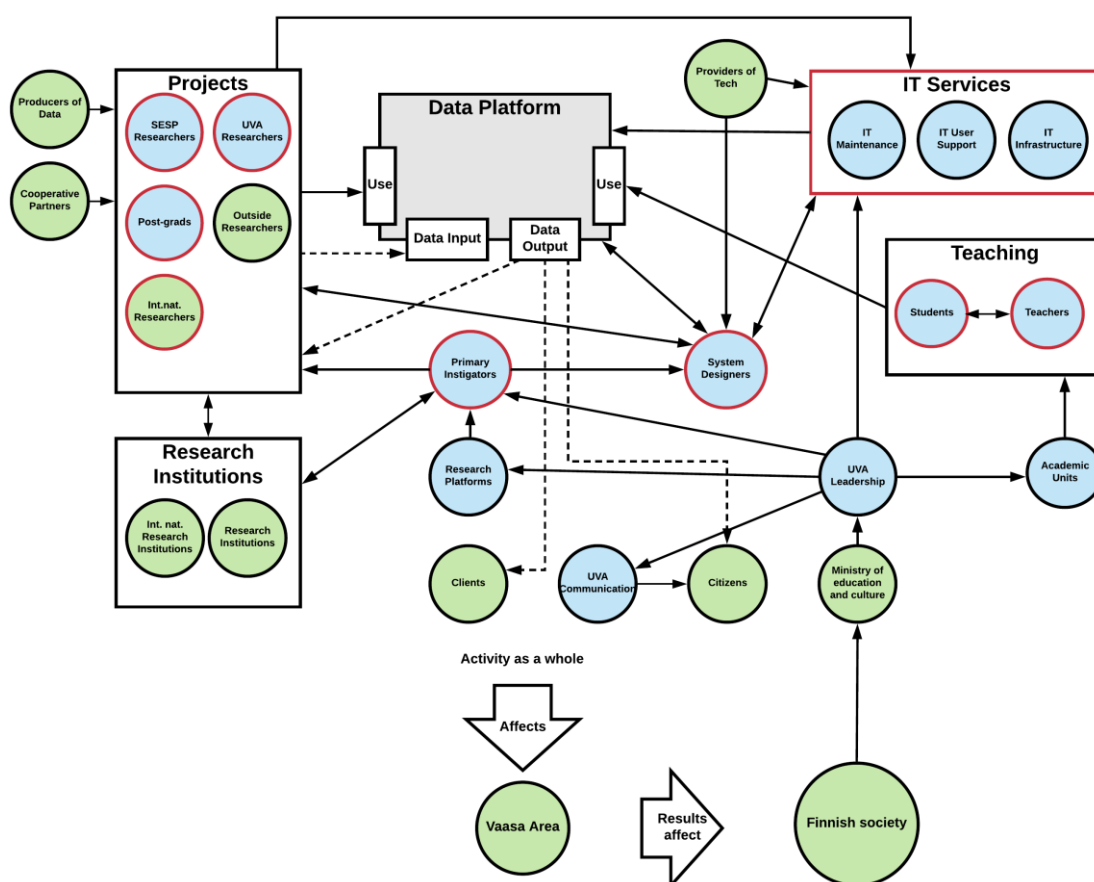


Figure 8. Results of stakeholder analysis.

These groups are more clearly presented in table 3. As Friedman et al. (2008: 88) suggest, in some designs very wide groups can be considered indirect stakeholders, some research decisions have to be made of deciding which groups are actually feasible and useful to include in conceptual studies. In this study, priority is given to groups that are strongly affected.

Table 3. List of stakeholders chosen for further analysis.

Stakeholder	Direct or Indirect
UVA Researcher	Direct
SESP Researcher	Direct
Int. Nat. Researcher	Direct
Doctoral Student	Direct
System Designers	Direct

Teachers	Direct
Students	Direct
Primary Instigators	Indirect
SESP Primary Instigators	Indirect
IT Services, maintenance	Direct
IT Services, user support	Direct
IT Services, infrastructure	Indirect

5.4 Phase III: Second technological investigation

As a result of both the first technological investigation and the second empirical investigation, it was clear that the first prototype would not be able to fulfill the organizational role envisioned. Furthermore, scaling up the principles of the first technological investigation by outside platform provider was not possible for project related reasons.

Scaling up with cloud services was also investigated tentatively by using personal Azure credits of the researcher and free test use provided by Google and Amazon. These platforms could have provided scalability but for the costs involved this avenue of approach had to be abandoned. Issues in the cloud-based approach concerning privacy, security, contractual usage of the potential data, knowledge building, and technical processes were left unsolved. Instead, another approach had to be found.

IT-Management of the university was identified by a co-designer as a possible provider of computational resources necessary and more importantly, it could be realistically approached with the zero budget the researcher had available at the time.

5.4.1 Securing of system resources

During September and October, initial negotiations with the IT-department were concluded. Access was provided on computing power and storage, based on a system consisting of a cluster of four Intel Xeon E5-2630 v3 based computers, each having 16 logical processors, 256 GB of memory and 19 TB of storage. Of the storage capacity, approximately 15 TB was promised as usable space for the prototype. As there were other uses for the computational resources, and to guarantee both the safety of the experiment for

the infrastructure and the flexible division of resources between nodes, cluster was built as virtual computers. Additionally, as VMware was already in use by IT, this approach was the most sensible and it provided a proper way to grant access to the resources for the researcher.

Several teething problems regarding the cluster were found, including the inability to upload data, remote access, and network functionality of the environment assigned. These were mostly related to the fact that providing access to these resources outside of the IT department was new. Eventually, these problems were corrected.

5.4.2 Design of the second prototype

Initially, computational resources were divided into ten machines, two would serve as master nodes and eight as slave nodes. This proved to be too much, and in order to reduce the overhead of storage space and gain more HDFS space, slave node amount was dropped to eight and the HDFS storage capacity of each was increased. To keep the storage use in the estimated range, after experience proved this to be necessary, all hard drives were thickly provisioned and eager zeroed in the deployment phase.

As security was identified in the first and second empirical investigation as one of the initial key values, alongside versatility and usability, the design decisions of the prototype were based on these. Additionally, three key principles identified by Begoli & Horey (2012) and discussed in more detail in chapter two, *supporting a variety of analysis methods*, *one size not fitting all* and *making data accessible* were included in the design and in the initial evolution plans.

From the Hadoop ecosystem and of the HDP package the user interface providing the usability necessary was identified as Zeppelin. Compared to command line interfaces of Spark, Beeline of Hive, or the submission of MapReduce jobs, the notepad approach provided a much more user-friendly way for the potential users to perform their first BD related research. Additionally, by offering only web-based access initially, a lot of the security-related matters could be investigated, tested and resolved in steps during the evolution of the platform. Compared to offering direct shell level access, this approach made the initial securement of the prototype more reliable, as Zeppelin supports also HTTPS and as a complementary measure, accessing the UI is only possible from the university network. Zeppelin also provides the initial versatility required, as it permits the use of

several different interpreters and allows the users to use the programming language for the analysis they are most comfortable with.

In figure 8, the resulting network design of the platform is illustrated. More in-depth technical description and the changes based on the principles is provided in chapter six. All cluster computers are connected via an internal private network. A router VM provides a point of access with several DNS addresses, to provide a user-friendly way to reach the web-based UIs. As the actual virtual machine providing the hosting of the Zeppelin UI has the Fully Qualified Domain Name (FQDN) set the same as the DNS name, the user can access working UI even if the actual hosting of the UI is in different VM that the DNS directs to.

For the security reasons, web-based UIs used in the administration of the cluster are not reachable from outside. Besides the VMs making up the cluster, there exists additional VM named Graphlan in the figure, running on Linux with minimal resources but with graphical desktop activated, to provide access for the administrative UIs. Access to this VM requires access to VMware and the university network.

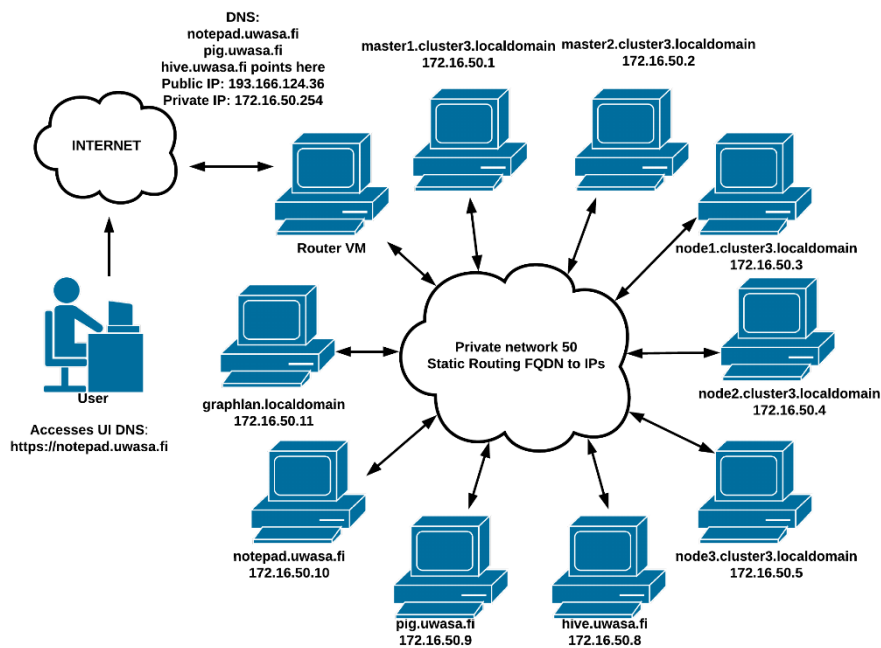


Figure 9. Network infrastructure design of the prototype 2.

5.5 Phase IV: First conceptual investigation

The first initial conceptual investigation was based on knowledge gained during the literary review, onboarding to the case, reviewing of the project documents, interactions and discussions with various project personnel, and during the drafting of the initial specifications document for the Big Data platform provider. This process as a whole is described as the first initial empirical investigation. This is congruent with the view of Friedman (2008: 72) where methods employed in the empirical investigation can include the entire range of quantitative and qualitative methods of social science review, including collection of relevant documents and interviews. Initial empirical investigation cannot be described as a rigorous activity or very structured, nonetheless, it provided valuable information as it uncovered initial values and some potential value conflicts. In the second empirical investigation, in the identification of the stakeholders, more initial values and possible conflicts were found, when the relationships between stakeholders were mapped. In this first conceptual investigation these initial values were explored and based on these, an empirical investigation was designed to reveal more hidden values and gain explicit knowledge on how stakeholders prioritized values in identified potential value conflicts. Several values were initially discovered during the earlier investigations. These are described in table 4. Most of these values are directly related to the properties of the platform, some to the organizational usage of the platform and only a few to the general human values.

Table 4. Initial identification of values.

Values related to platform	Organizational Context	Values in Life
Affordability	Co-operation	Openness
Connectivity	Distributability, Shareability	Privacy
Storage of data	Interestingness of results	-
Developability	Research	-
Usability	Teaching	-
Security	-	-
Versatility	-	-

5.5.1 Identification of initial key values and value conflicts

To enable the development of design principles to last the evolution cycles of the platform, conflicts between these initial values must be solved. Interviews would have to be able to guide in the prioritization. Additionally, expectations were that after interviews, more in depth-values and additional information would be gained by analyzation of the results both qualitatively and quantitatively. Initial value conflicts identified are presented in table 5.

Table 5. Initial value conflicts.

Conflicting value	Conflicting value
Openness	Privacy
Distributability, shareability	Security
Usability	Versatility
Storage of data, connectivity	Affordability
Research	Teaching

5.5.2 Design of the interviews

The interview was designed around themes with one of the parts consisting of a quantitative measurement of the relative importance of certain values. One of the reasons for choosing the approach was the critique of VSD by Manders-Huits (2011: 278–279). She states that empirical methods in the form of a survey in VSD are questionable for two reasons.

Firstly, modern technology as a whole and the limitations and possibilities it provides, especially specialized technology, is not understood widely. People may have mistaken beliefs about the technology or issues concerning it. Secondly, it is not always clear what the stakeholders actually mean when they are mentioning certain values. Meaning can change due to how well values are defined in the research and how the values are interpreted and experienced by the stakeholders. (Manders-Huits 2011: 278).

Thus Manders-Huits (2011: 279) claims that interviewing stakeholders can be in loose grounds, if not giving “ground or substance to values”. Usefulness or validity of the results can be questioned if the values are too abstract and multi-interpretable. Additionally, Manders-Huits (2011: 279) points out another problem in the empirical part of VSD – stakeholder opinions are considered having a shared and fixed point of view. Opinions and values of people can fluctuate based on new experiences, knowledge, and insights. Additionally there exist numerous interpretations of certain values and normative positions. Therefore Manders-Huits (2011: 279) would recommend a more deliberative method than a survey to properly identify the issues. (Manders-Huits 2011: 279).

The design of the interview tries to take all this critique into the account. Firstly, the first three themes are chosen to describe very concrete actions, use cases and harms. This should avoid the first reason for critique by Manders-Huits (2011), lack of understanding of the technology affecting value prioritizations. If participants conception of related technology is indeed mistaken somehow, it is evident in the actual use case they describe and can be noted. The second point of critique, the inaccuracy of values, the difference in meaning of each value to each participant is attended in the quantitative fourth theme, where while prioritizing each value, the participants also tell what they actually mean with that value.

According to Tiainen (2014: 3–5, 16–17), a themed interview should be based on theoretical framework if the approach is what Deetz (1996: 198) describes as dissensus based with a priori formation of the frame of reference. In this research, the frame of reference is described based on the previous empirical, conceptual and technological investigations. Especially important was the second empirical investigation where the context of the platform is envisioned in relation to organizational context, without any temporal suppositions. In other words, the result of the first empirical investigation does not take into account *when* the system is described. The organizational context for the system is the same, no matter if we are discussing the system in the early development phase or in the later stages. The frame of reference is illustrated in Figure 10.

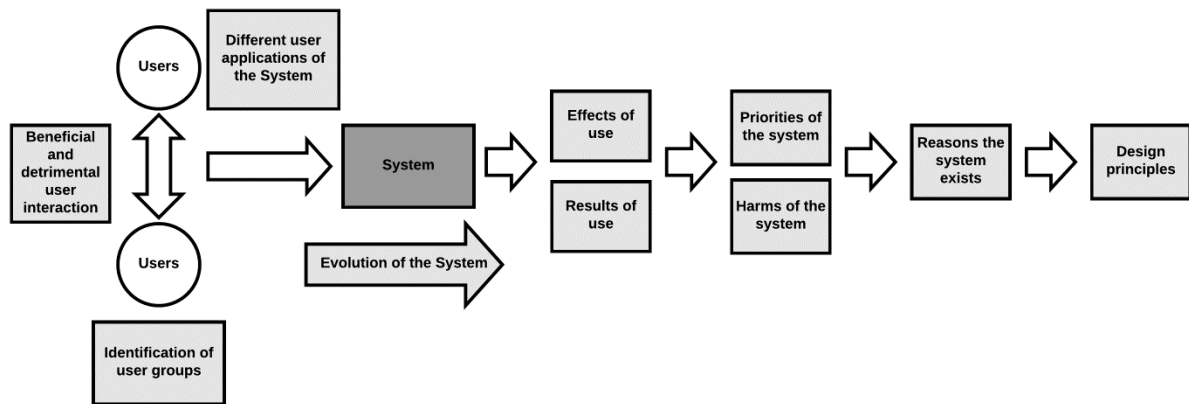


Figure 10. The frame of reference for the interviews.

Based on this frame of reference, the interview was built on four themes: a system in entirety and the lifecycle of the system, usage of the system and the users, personal usage of the system, and quantitative prioritization of recognized values regarding the system. In the design of the actual questions, open-ended questions where the answer could not be just “yes” or “no” were preferred, a principle proposed by for example Tiainen (2014: 17).

The first theme was concerned mostly with the temporal existence of the system, the key axis that the design principles would affect. Additionally, the negative effects of the system, detrimental user actions, and disadvantages during the lifetime of the system were probed. First theme, system in entirety and the lifecycle of the system consisted of the following questions:

1. How do you see the entirety of the system?
2. How do you see the lifecycle of the system? Where does it start, what happens during the lifecycle and how does it end?
3. What kind of harms or disadvantages related to the system can you imagine? These can be related to direct usage of the system, results of usage, relations or conflicts between different users, user groups or related entities?

These three questions were beforehand expected to provide insights to mainly to how the interviewees saw the evolution of the system, reveal harms related to the effects of use,

results of use and detrimental user interactions. As they were open-ended, there were expectations that other relevant and interesting views would arise. Q1 was designed to get the interviewees to talk about the system freely, how they see the platform, in what kind of context, what kind of interactions between organizational entities and user groups related to the system exist, what kind of results the system provides and thus hopefully reveal inherent value prioritization in that view. Q2 was addressing the ideas and visions held by the interviewees of the system evolution directly. Q3 was expected to provide insights about negative value generated either by system directly just by existing in the organizational context, the act of using it for different purposes or by results the usage of the system provides. It also attempted to cover possible detrimental effects to user or user group relationships any of the previous could generate.

The second theme, using the system and the users, consisted of three questions:

4. What the system should be used for?
5. Who should use the system?
6. What use would you consider the most important?

This theme was interested in gathering insights for the reasons the system exists, identification of users and user groups and prioritization of the use cases. Q4 was designed to reveal indirectly what the interviewee saw as primary reasons for the existence of the platform with Q6 repeating the same avenue of approach directly and more explicitly, after the interviewees had thought and pondered the platform from the point of view of the users and user groups in Q5.

The third theme, personal use of the system, was mainly directed at the interviewees who would directly use the platform to gather more technical level data for the initial technical design and prototype starting points. It was expected that this theme would provide insights for direct different user applications of the system, both direct and indirect such as educational or promotional usage, and provide insights indirectly for the priorities of the system and effects and results of use. Additionally, for interviewees in more managerial roles, expectations were that answers would reveal indirect usage with different organizational entities. Questions in this theme were as follows:

7. Would you personally make use of the system in some way? Direct or indirect? If so, how? If not, why not?

8. If you would like to make use of the system in some way, what would be the most substantive in your use and what would you consider most important in it? Perhaps something special or different?
 - a. Regarding your data? In the format or amount of data? Velocity of the data? Legal or contractual restrictions regarding data? Ownership of the data?
 - b. In the analysis and processing phase. Application of for example machine learning? Qualitative or quantitative techniques?
 - c. Regarding results? Ownership of results and sharing of them? Commercial usage? Results reached by different means of co-operation?
 - d. Other direct or indirect utilization of the system. What would you consider central or essential in other utilization of the system?

Q7 was designed to be open-ended enough to be able to include ideas regarding personal use of all kinds and also the indirect utilization of the platform in the managerial and organizational role of some of the interviewees. Q8 included several sub-questions, aimed at discovering technical pre-requirements of the possible uses, results, and effects of the usage but also included open-ended subjective prioritization prompts. The last sub-question was designed to prompt the interviewees to think again other possible personal, direct or indirect, usage of the system.

The fourth theme was a quantitative part of the interview. As was mentioned, the design of the fourth theme was based on the mentioned critique by Manders-Huits (2011) and the awareness that in conventional content analysis, comparison of prioritizations could be difficult with only qualitative material. Hsieh (2005: 1281) describes that results of the conventional content analysis would be limited to concept development or model building at most, unlike techniques such as grounded theory method or phenomenology. It was presumed that the limited quantitative part of the interview would provide additional reinforcement for the qualitative material. As the initial key values and possible value conflicts related to the platform were already identified, they were used in this theme. Interviewees were asked to prioritize them by choosing five most important to them and to number them from one to five, one being the most important concept to them in the context of the platform, second the second-most important concept and so on.

In this theme the values related to the properties of the platform, organizational context and universal values were mixed. The act of choosing five central concepts amongst several, some options among them lower level concepts than others, was meant to force the participants to actively think prioritization conflicts that would most likely to be encountered during the lifetime of the platform. Picking out a lower level concept would suggest a quite high personal priority for the concept in question. Values of co-operation and interestingness of results were divided into subcomponents, in order for the results to offer more detailed guidance on the prioritization.

Concepts were as follows, ordered alphabetically in the form in both languages. As the interview form was created originally in Finnish, then translated to English, concepts were in a different order on both forms. Additionally, there was space provided if the participant wanted to add a new concept or value.

- Affordability
- Connectivity
- Co-operation, Scientific
- Co-operation, Commercial
- Co-operation, International
- Co-operation, Local
- Data Storage
- Developability
- Distributability, Shareability
- Interestingness of results, Scientific
- Interestingness of results, Commercial
- Openness
- Privacy
- Research
- Security
- Teaching
- Understandability
- Usability
- Versatility

As these concepts were created before the interview, there existed a chance that participants could consider concepts or values important to the nature of the platform, values or

concepts that the researcher had not thought of, if only premediated options were offered. On the other hand, new values added by the participants would be expected to be unique and remain on low ranks in the quantitative analysis. I decided that despite this, offering an opportunity to add a value would be included, as it would secure against me missing a central value and in that case, a significant amount of related answers would disclose it.

5.6 Phase V: Third empirical investigation

In the third empirical investigation, the interviews designed in the previous section, second conceptual investigation, were actually conducted and the results analyzed.

5.6.1 Conduction of the interviews

Interviews were estimated to take 30 minutes and with each participant, a suitable time was arranged. As stakeholder groups were identified, potential participants were approached. If other contact methods proved unsuccessful, a suitable date was arranged by approaching potential participants in person. All initially identified participants could be interviewed. Interviews were conducted during a time period of 3.10.2018 – 2.11.2018. Most interviews took place in the separate offices of the participants during their work days, exceptions being five interviews conducted in reserved meeting rooms, one interview in the home of the participant and one interview in the recreational room of a student organization. No compensation was offered for most of the interviewees, only student priced meal for the two student participants. Estimation of the length of the interviews proved relatively accurate, the shortest interview took 18 minutes and longest one hour and 20 minutes, with the rest of the interviews staying in the 25–40-minute range. Interviews were recorded and then transcribed manually, resulting corpus consisting of approximately 30,000 words.

Participants were not offered any preparatory information before the interview. As the case project had already been running for almost two years it was anticipated that consensus voice representing the original aims and goals of the project would be strengthened in that case. Instead, it was decided to combine both the warmup Tiainen (2014: 15) recommends and an introduction of a new view of the platform (Appendix 4). This was

hoped to get participants already familiar with SESP-project to view it from a new direction and the participants unfamiliar with the platform to achieve an understanding of the context of the interview.

Additionally, warmup procedures and sensitivities recommended by Tiainen (2014: 15) were observed by asking the date of birth instead of age and participants were allowed to describe their occupation in the detail they chose. Participants were prompted to read the introduction on the forms (Appendices 1 & 2) and in the context of the meta-level picture (Appendix 4) I read the warmup questions without expecting or asking for an answer. Their only purpose was an attempt to get participants oriented in the context of the interview. A test interview was performed and due to that, one participant was interviewed twice.

Based on the availability of representatives of previously chosen groups for interviews and ensuring proper coverage, the final list of survey participants was as is represented in table 6.

Table 6. Survey participants.

Code	Stakeholder Group(s)	Occupation
I1	Student, Bachelor	Bachelor Student
I2	IT Services, Manager	IT, Manager
I3	SESP PI, PI, UVA Teacher, UVA Researcher	Professor
I4	SESP PI, PI, Outside Teacher, Outside Researcher	Professor
I5	UVA Researcher, SESP Researcher, IN Researcher	Assistant Professor
I6	UVA Researcher, SESP Researcher, IN Researcher	Assistant Professor
I7	SESP Researcher, UVA Researcher, UVA Teacher	University Lecturer
I8	UVA Researcher, Doctoral Student	Grant Funded Researcher
I9	IT Services, Infrastructure, Security	IT, Information Security Manager
I10	SESP PI, PI, UVA Teacher, UVA Researcher	Professor
I11	UVA Teacher, UVA Researcher	University Teacher
I12	SESP PI, PI, UVA Teacher, UVA Researcher	Professor
I13	SESP Researcher, UVA Researcher, Doctoral Student	Project Researcher
I14	UVA Teacher	University Teacher
I15	Student, Masters	Master Student
I16	SESP PI, PI, UVA Teacher, UVA Researcher	Professor

Many interviewees represent more than one stakeholder group as was to be expected as the context was a research project in a small university. Professionally, all the relevant university roles were represented from professors to bachelor student. Support services of teaching and research were represented by IT, with the head of IT and more technical view provided by the information security manager. All the stakeholder groups identified in table 3 were covered by participants. The oldest participant was born in 1959, youngest in 1992 while the median of the years of birth was 1975. Participants chosen could have had a better representation of genders, as all participants were male. On one part, this limitation can be seen as a result of gender over presentation in primary instigator roles, especially in SESP-project and in technical fields in general, but it would have been possible to address this with a better sampling of interviewees. Researchers, teachers, and students would have been available. However, while recognizing this limitation, I would expect the effect of this limitation to be negligible. With different sampling, results would be expected to differ, but most of that would be due to the change in the ideas and values of the new interviewee, not because of their gender.

Participants consisted of 13 people speaking Finnish as their first language, one having Swedish as their first language, and with two participants English was used as a common language. Therefore two sets of interviews were created, one in Finnish and one in English. These can be found respectively in appendices 2 and 3. The survey form was designed originally in Finnish, then translated to English. Of the interviews, 14 were conducted in Finnish and two in English.

5.6.2 Interview results

The qualitative part of the interview was analyzed with conventional content analysis method. Hsieh & Shannon (2005: 1279) propose that with this approach the interview questions should be open-ended as was used in the study. They see the analysis of the data starting with gaining an understanding of the material as a whole by reading it repeatedly. Then data is read word by word to derive initial coding. Then the data is approached by the researcher making notes of their impressions and create the initial analysis. By iterative recoding, coding labels should emerge that are reflective of more than one thought. These, Hsieh et al. (2005) see to raise directly out of the text becoming the new initial coding scheme. Categories are then established based on how the codes are

linked and related. With these emergent categories, meaningful clusters are established. To ensure broad enough clusters, to contain large enough amount of codes, suggested amount of clusters is between 10 or 15. (Hsieh & Shannon 2005: 1279).

Interviews were transcribed and imported into NVivo program where the iterative analysis took place. Interviews resulted in transcripts containing 30,000 words. As these interviews were planned, performed and transcribed by the same person it was expected the gaining understanding of the data as a whole would be straightforward. This turned out not to be the case, but eventually coding consisted of 466 marked passages to different codes, with many passages of text belonging to multiple codes. Eventually, nine clusters emerged with some containing sub-clusters. These are listed in table 7.

Table 7. Emerged clusters from the interview analysis.

Cluster	Containing
Values	Human values related to the purpose of the platform
Evolution, Strategy, Development	Important aspects by the participants on the development of the platform
Challenges	Platform related challenges, development related
Harms	Negative effects related to the platform
Users and usage	Objectives of use, possible forms of use, identified user groups
Use cases	Concrete low-level ideas of utilization of the platform
Data sources	Concrete and more general ideas of data sources that could be utilized with the platform
Potential benefits	Benefits or positive value related to the usage of the platform
The lifecycle of the Platform	Ideas related to the start or the end of the platform

5.6.3 Harms related to stakeholders

Friedman et al. (2008: 88) describe that in VSD identification of benefits and harms for each stakeholder group should occur right after the identification of stakeholders. One suggestion of theirs is using personas, but they do not close out any method. The strongest

guiding word they give is “systematically”. At least in this study, mapping of benefits and harms via imagination proved unsatisfactory results. As a consequence, I decided to execute the third empirical investigation, the interviews to explore potential benefits and harms the system and the usage of it could entail, and only afterward investigate the harms and benefits based on empirical data. These are clustered in Harms and in Potential Benefits, as represented in table 6. These two clusters were then further processed visually in LucidChart to create the new relationships and discover the containing classes and the sub-categories within. If the emerged clusters would have been defined with more detailed build-up of codes, this phase would have been unnecessary. However, as Rowley (2010: 267–269) describes, there is not necessarily only one way to proceed.

Interpretation of the new relationships results to following largest classes of Harms cluster, is represented in table 8. In addition, several orphaned classes emerge, including privacy concerns related to the idea of continuous integrative nature of large data storages, and what kind of effects that direction of development could have in the future for people globally.

Table 8. Classes and categories within the Harms cluster.

Harm Class	Containing categories
Control of the Platform	Information Security, Malevolent usage effects and reasons, equality in the distribution of system resources
Usability	The ability of users to use the system effectively, additional learning, usage of system autonomously of the support, versatility
Lack of Know-How	Education usage of the platform, user adoption challenges from lack of knowledge of operations, incorrect expectations, resistance, unknown area
Material related	Data ownership related issues including ownership of data collected by devices, ethical, contractual and juridical concerns, shareability of the collected data, data collected covering one research area, data accuracy

In the class *Control of the platform*, malevolent usage was related to three answers. The humanity of the users was mentioned and interpreted to mean that human beings have a distinctly different understanding of the outside environment from themselves, the world of perceptions, consciousness and mental states that Iivari (2007: 42) mentions. This could result in varied actions, some perhaps more related to destroying value than creating it, or hindering the purposes of the platform instead of advancing them. The performer of these actions would perhaps not see the results of these actions as negative due to the reasons discussed, requiring enforcing of actions to be available in the platform. In this malevolent usage category, appropriation of system resources and usages of proprietary data for commercial benefit or potential hacking purposes was identified by two participants. Division of system resources was also mentioned in this class from the point of view of equality and results of the failure of the anonymization of the data was also raised by a participant from the point of view of privacy.

Answers in the class of *Usability* contained concerns of the participants related to the usability of the platform. Usability has several different understandings in the public domain. Research directly focusing on usability depends usually on different heuristics to evaluate it. In this study, participant concerns in this class were interpreted to belong to three different categories of usability. Firstly, very concrete examples of usability in the interviews were related to the harmful effects of poor usability on the efficiency of the users. This was interpreted to be understood in the context of the efficiency of the users in their work in the current academic environment with tight budgets and increasing emphasis on different ways to measure and report the personal efficiency and performance. Poor usability of the platform, including the technical stability and the organizational process of usage, makes their work progress slower, which reduces their efficiency.

The second category interpreted in the class of *Usability* was the usability in the sense that poor usability requires additional learning. This had two aspects. One aspect was resource related with one perspective discussing how to secure the resources necessary for the user education and tutoring, and another perspective related to the reluctance of the users to participate and do additional learning to gain enough competency in the usage of the system. This latter perspective could be related to the efficiency concerns discussed earlier. The third category interpreted from the Usability was related to the autonomy of the users. Usability meant that the platform could be used autonomously, without the necessity of support personnel of various capacities being a constant part of the process. It is understood in the sense that co-operative approach of the usage of the platform could

be imagined to be less efficient, to require more planning and more rigid form of operations, and to be in some degree incompatible with the idea of academic freedom of a researcher. Even if this conjecture of the whys and wherefores of autonomy being highly valued in the context of usability is incorrect, the autonomy of the users was clearly valued.

Class of *Lack of Know-How* contained several categories. The first harmful effect was directly related to the potential education aspect of the platform. What the platform teaches for the students in their various phases of education is directly related to the architecture and the developed usage process of the platform. If they do not match the best practices of the industry, the effect of teaching incorrect operations and operations environment could be harmful. One participant was especially concerned with effect, and the interpretation behind the concern is how the participant sees the value of education. It could be understood that the value of education is here most related to the provision of the necessary skills to sustain themselves.

The second category in the class was related to the lack of know-how affecting the user adaptation of the system. Lacking the necessary knowledge and skills in usage of the platform has various degrees. If the skill gap between the actual skills of the user and the skill level the system usage requires is wide enough, it can make autonomous usage of the platform impossible. Secondly, if the skill gap is narrower, even then the adoption will not be perfect, ie. the platform will not be used in the best possible way and/or all the potential results from the data of the researcher will not be discovered, if the knowledge of the possible analytical approaches the platform enables is not understood.

Third harmful effect recognized in the class of *Lack of Know-How* was incorrect expectations of the platform due to the lack of knowledge. One participant describes this lack of actual knowledge of the Big Data and especially of the actual implementation of the analytic and storage platforms resulting in thinking where the system is thought of as a black box. Interpretation of this is that the general interest and discussion surrounding Big Data, and the somewhat related concept of IoT and of the result of Big Data analysis method, artificial intelligence, generate expectations that are disparate from the resources provided for the implementations. Data goes into a black box, great results emerge. Easy. A participant described this as the colliding of the hype and the real world, resulting in bewilderment and resistance. Resistance was recognized as the fourth harmful effect from the lack of know-how, independently of the incorrect expectations by one

participant. Here, it was described resulting from the difference of the system compared to previous systems and resistance itself emerging from the time and effort the learning requires. Again, it would be tentative to interpret this with the context of efficiency and efficiency expectations.

The fourth harmful effect of lack of know-how was categorized as unknown area. Here a participant, with a strong background of IT systems and deep understanding of the current challenges and opportunities in the field, described the general lack of know-how resulting from the fact that the field of BD analysis and utilization is generally uncharted territory. There does not exist one correct view of the matter, as there does not exist an established mode of operations or of the structures of the system. This view is congruent with the views presented in the relevant literature. The result from this lack of know-how is interpreted to be research needs in developing, maintaining and operating the related technology and in leading such activities.

In the *Material Related* class of harms, the first effect consisted of issues related to data ownership. This was referred to by six participants. Five of the references were describing the harms related to the utilization of data covered by various restrictions. One interpretation of the harm is understanding it as a limitation on the potential of the data, therefore preventing reaching of the results the data had the potential to provide if it could be analyzed and combined without any restrictions. Other interpretation of the harm is related to the cause of the restrictions. If there were no restrictions based on data ownership, ethical issues including privacy-related concerns, legal issues as GDPR related concerns, or contractual issues with private company provided data, usage of the data could generate concrete harms for larger groups. These harms can occur even if the restrictions are existing and enforced, but can be bypassed. Users have to trust the real efficiency and effect of these restrictions, in order for them to trust the platform enough to provide data for it. Related is the data ownership issue of these restrictions constraining the shareability of the data.

Moreover, the data ownership issues exist related to the ownership of data collected by devices, for example, the data collected by the electricity consumption meters. An electrical company in question owns the meter but is not certain if the data the meter collects is legally considered as data of the customer, or is it owned by the company.

In a material related class of harms were additional observations by participants how the inaccuracy of the collected data and the concentration of data sources focusing on one research area could be harmful. The former was interpreted to mean the inaccuracy of the data leading to incorrect conclusions with resulting harms affecting both the researcher, project and the partners utilizing the data. Possible data sources utilized by the platform concentrating on one area of research only, clearly leads to narrowing of the utilization of the platform, which is in contradiction with the idea of a versatile platform.

5.6.4 Benefits related to stakeholders

The emergent cluster of Potential Benefits was further processed similarly to Harms in the previous chapter, forming classes visually in Lucid chart. Resulting classes are described in table 9.

Table 9. Classes inside the Potential Benefits cluster.

Benefit Class	Containing
Based on Data	Benefits from storing the data, sharing the data, specialized data availability
Education	Benefits related to teaching and learning
Facilitator	Benefits for the stakeholders by the facilitation aspect of the platform
Co-operation	Benefits for the stakeholder from the co-operation related to the platform

First beneficial effects on class *Based on the data* are related to the benefits provided by the act of storing various data. Storage of the data continuously and performing this for longer time periods would result in potentially unique research opportunities as described by one participant. This, in turn, would result in benefits for several of the stakeholder groups, in and outside of the university. Benefits resulting from unique research are self-evident for research related stakeholders and clear for the research organizations, but the possible indirect side effects for the indirect stakeholder entities such as the nearby municipalities, communities, and business much harder to accurately depict. They could be hypothesized to be net positive but the scale of them could be insignificant.

Second beneficial effects depicted in the class by one participant are benefits related to the sharing of the data. Commercialization of the collected data was reinforced by views of a second participant. Additionally, benefits for the stakeholders from the sharing of the data include the opening of the data for use of the external entities non-commercially. Benefits from this can be interpreted to include involvement and interest of the surrounding community in the research. Benefits of such interpretation are manifold for the stakeholders and in-line with the benefits championed by the citizen research, including effects from research being more accessible and as more understood part of the society. Another view presented by a participant in the sharing of the data involved marketing aspects of the university, research, and a specifically mentioned research project. It was envisioned by a participant that the marketing communication of the university could make use of shared data for a more visualized and concrete way to present research done.

Availability of the specialized data was the third beneficial effect in the data related class. With it, the participant meant that especially in his area of research, data necessary for research is not publicly available. It has to be gained by trial and error, resulting in difficulties in estimating the time required for each action and can be interpreted to lead to the same efficiency concerns discussed previously. If this proprietary data would be available in the platform or via the projects it facilitates, originating from the partners, then the trial and error phase would not be necessary. There would be benefits for both the academic side of the process and for the industry side of the process.

Second benefit class was *Education*, consisting of benefits on both sides of the equation, both for the students and educators. For the students benefit identified by one participant included the learning of the right methods, understanding of them and seeing in the results that they produce qualified data. Two participants saw some possible benefits in the usage of the platform in relation to bachelor theses, more so at the level of masters and one participant saw benefits in the doctoral level studies. For the educators, benefits were seen by one participant, mostly related to the concept of the platform existing as a source of materials.

The third class of benefits for stakeholders was *Facilitation*. Three participants identified the platform working in the role of facilitation by their interview answers. Firstly, one participant saw the role of the platform as providing the benefit of the direction of energy production and consumption in the Smart Grid by enabling estimation of production and consumption based on the data stored. Related is the second beneficial effect, facilitation

of the energy transformation to the production of renewable energy sources via usage of the platform provided analytics for the direction and guidance of the necessary measures. Third beneficial effect for the stakeholders in the class was the development of new business models via analyzation of the customer behavior that the platform enables. Facilitation benefits affect several stakeholders. First and second effect are related and can be interpreted to be related to the values of continuity of the human race and environmentalism, or with a more critical view of seeing these values as a tool to enable the pursuit of the real value and objective, conduction of research and acquiring new knowledge.

Fourth and last category of beneficial effects is described by *Co-operation*. It was seen, with three aspects, with one participant expressing each aspect in the interviews, as a connective factor, a system integrating IT services with the research, and by providing benefits for humanities. With connective factor participant meant that it would connect the different laboratories existing and planned, by providing a common platform for storage of data and analysis of the data. Furthermore, by connecting these separate entities it would also include education and teaching more directly in the research.

By existing the system would integrate IT services more tightly with the researchers, a role for IT services strongly advocated by the participant. Benefits would include increased lower level co-operation resulting in a more efficient distribution of experiences, knowledge, and methods as the openness and overt sharing were identified as one of the strengths of the university community, both in the support services and elsewhere. Lastly, by providing benefits for the humanities, the platform would create lower level co-operation cross the disciplines in the university as the platform should be versatile enough to support a wide range of analysis and research methods.

5.6.5 Quantitative value prioritization by stakeholders

Participants of the interview asked to prioritize concepts and values they consider essential to the platform in theme 4. As participants prioritized the values marking with one the most important in the context of the platform, with two the second most important one, these were analyzed simply by giving the first priority five points, the second one four points until all five prioritized concepts were processed. The full table is available in Appendix 5. In figure 11 top results are presented, cut-off being at 10 points.

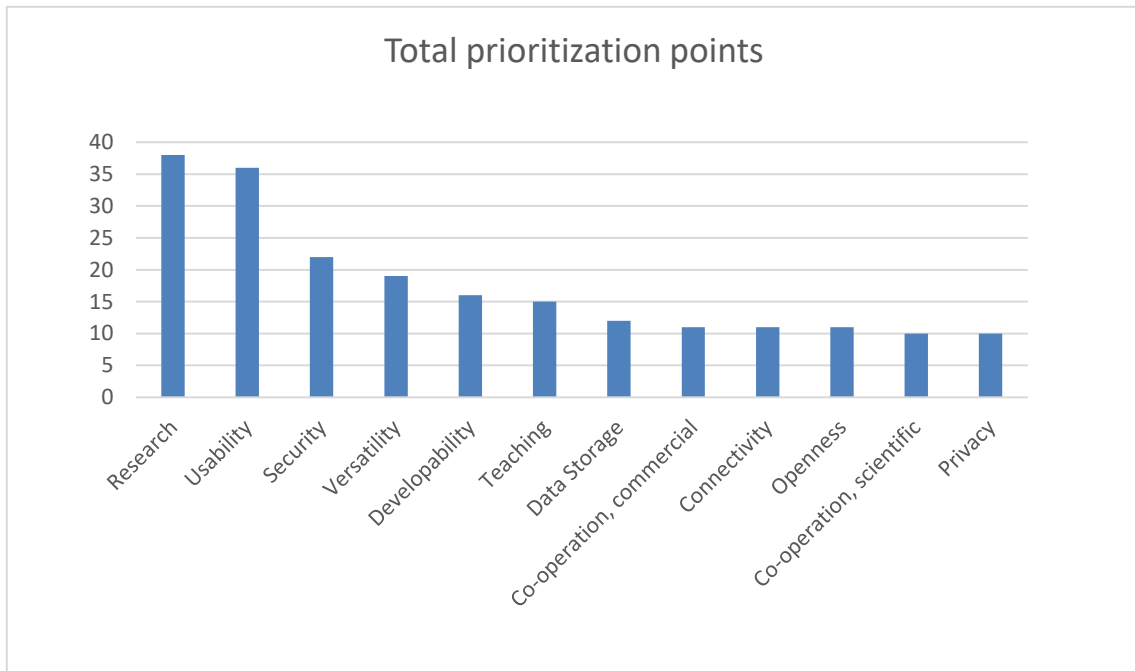


Figure 11. Highest prioritization concepts by points.

It is not possible to compare all the concepts directly from the general results, as these differ in scale and in nature. What can be inferred is that in general, research is the most important concept regarding the platform for the participants as a group. Values and concepts are difficult to compare, as the exact meaning given to each value differs by each participant.

As participants were asked to discuss and explain what they saw as the meaning of each value, as they answered them four questions, it is possible to categorize the meaning of research to several main categories. Firstly, and the main reason in the context of prioritization, it is used directly in the sense of *reason of existence* for the platform as described in the frame of reference. Further, it is used in the sense that research *is* why university exists, therefore the platform has to be related to research.

Following quotes are examples of the first category:

“Important platform for future research projects”. (I10).

“When we have a bunch of data we need to use it based on the research”. (I5).

“So that it generally can be used for research”. (I7).

In these the *research* the platform facilitates is why it should be developed and why it should exist. Why research itself is important contains various views, which are discussed more in depth in section 5.6.4. They include the ideas of gathering and developing knowledge, financial reasons and different views on the benefits of products of research. The platform is straightforwardly thought as a tool for purpose of research.

Research as a concept can be compared to teaching, in the importance of the duties of a university. It is clear that research is considered by the participants as a higher priority for the platform than teaching, as it has gathered over one and a half times the prioritization points compared to teaching. It is not surprising, considering the participants as a whole had more work-related interest in research than in teaching. However, appraising the interviews as a whole, there was a tendency for primarily teaching personnel and even students to some degree, to value research aspect of the platform over teaching, but this is harder to quantify exactly. The research was also considered the most important concept in the context of the platform by five participants, while usability was most highly prioritized by four participants. Amount of considerations as the most important for concepts is described in table 10.

Table 10. Amount of highest prioritization.

Value	Amount of highest prioritizations
Research	5
Usability	4
Co-operation, all forms	2
Connectivity	1
Interestingness of results, Scientific	1
Openness	1
Security	1
Versatility	1

As a conscious decision concept of co-operation and interestingness of results were divided into sub-concepts, to gain insights into their relative importance. This makes the relevant comparison of total prioritization points inadequate between all the concepts. In figure 12 co-operation and interestingness of results are summed. This brings value prioritization more in line with the qualitative results, where the co-operation was identified as one of the central values regarding the platform.

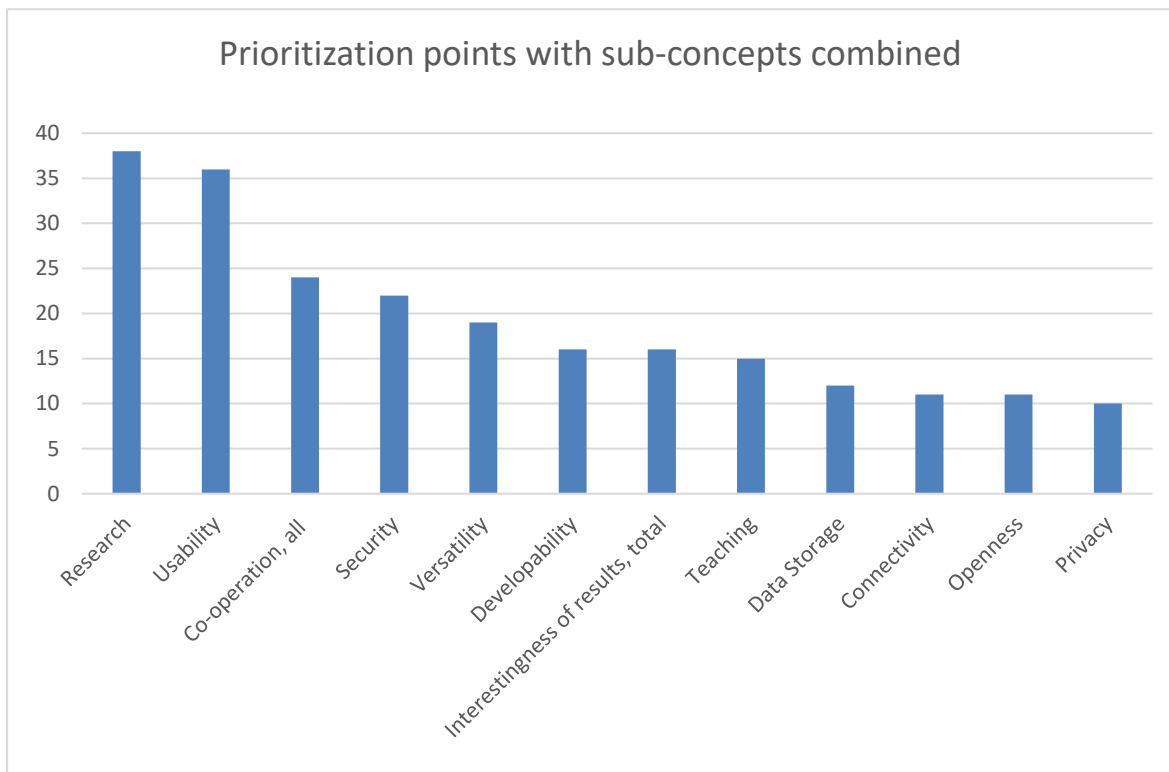


Figure 12. Co-operation and interestingness of results combined from sub-concepts.

Of the sub-concepts of co-operation, two were significantly prioritized, commercial co-operation and scientific co-operation. These were considered nearly equally important in prioritization points and both received one nomination as the most important concept related to the platform. These were described by the participants:

“Should be placed as the number one (priority) in order to get things happening with the platform. It should be prioritized as number one as it facilitates continuity”. (I1 describing commercial co-operation).

“If the platform takes off the ground, it will automatically generate contacts, know-how, networks and other things for us”. (I2 describing scientific co-operation).

Interestingness of results raises significantly in ranking once sub-concepts are combined into base concepts. It is directly related to the *reasons of existence of the platform*, if the utilitarian view is adopted and the platform is viewed as a tool for reaching certain results, which interpretation is supported by the descriptions:

“It is because of the partners we do this work, the platform must be useful for them”. (I13 describing interestingness of results, commercial).

“If the results are interesting, they are also important”. (I3 describing interestingness of results, scientific).

Insights can also be gained if higher level concepts are removed and only concepts describing platform on the more concrete level are left. Prioritization ranking of higher level concepts supports qualitative analysis by clearly depicting the felt relative importance of the concepts but offer little practical guidance on the priorities of lower level design decisions. Ranking of concepts related to the properties of the platform can bridge that gap. This is depicted in figure 13.

Usability is the most important concern for the participants. Some of the usability concerns are related to the concerns of user adoption such as:

“Bottom line on the usability – if it is too hard to use, people will not be using it, they do not want to use it”. (I8).

“Nothing works if it is not easy to use. People give up if the data is not usable”. (I14).

“Using of many systems may have ended even though the system is great, it is too hard to use”. (I13).

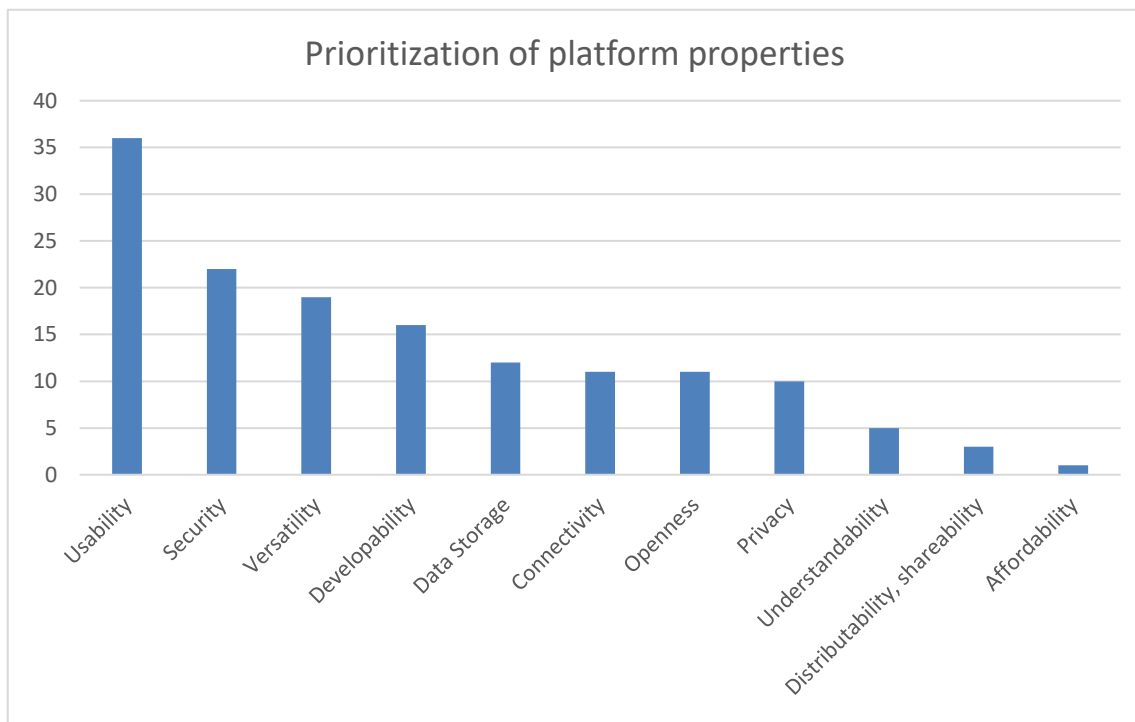


Figure 13. Prioritization of the platform properties.

Among other aspects of the usability was efficiency related meanings that can be interpreted to be related to the reason of the existence of the platform in the frame of reference – saving time compared to the situation where platform does not exist. Additionally, usability had resource limitation related concerns related to the needs of educating the users.

Participants prioritized security as the second most important property of the platform and interestingly only one participant saw it as the most important concept regarding the platform. It is an interesting result when considering the platform would hold massive amounts of data, possibly consisting of sensitive information subject to legal and contractual restrictions. Participants who prioritized security seemed to have the holistic meaning of cybersecurity, encompassing issues resulting from poorly conducted authorization and authentication, as the following examples show:

“There are activities done based on people’s information, so keeping it safe would be swell”. (I7).

“Different rights of ownership and access of the data, security overall”. (I16).

“I mean security of the data, security overall, if we have this kind of system containing confidential data, that nobody who should not be able to access it, should not be able to”. (I8).

Of the next high ranking properties, versatility was discussed on the sense university containing multiple different disciplines each with different traditions of research and to the variety in data existing in general. With this prioritization, they pointed out that the platform should be able to be used in several different ways.

With developability prioritization, participants meant that a system should be built in a way that system evolution would be possible. Some had a more general view of this, such as:

“It is related to connectivity, that it is not a project that ends in point X, but it is something that can be developed further in the future”. (I13).

While some had more concrete visions of the future:

“At the start, we do not lock our approach and we won’t ruin our chances to attach company infrastructure or living labs”. (I10).

5.7 Phase VI: Fourth empirical investigation

In fourth empirical investigation, a workshop was organized in order to gain a better understanding of how a larger group of potential stakeholders understand the function of the platform – how it relates to different organizations, organizational entities, and even individuals. It was also expected that the workshop would reveal directly and indirectly how the participants saw the function and purpose of the platform. This was expected to gain reinforcing input for the values identified and in addition, to provide insights into the planning of the organizational usage of the platform. This workshop was originally an idea of the co-designer of the platform and was planned and executed in co-operation.

5.7.1 Workshop

For the purpose of gaining more insights regarding the Big Data related opportunities in research, education and about internal or external collaboration a Workshop was organized. The objective was to merge ideas, insights, and thoughts of the participants into an information map about Big Data, potential actors and their relations. It was intended that the resulting information map would represent viewpoints into Big Data platform from multiple different disciplines and roles.

The workshop was held on 15.11.2018 at the premises of the University of Vaasa. Invites were sent out approximately one month and a half before the planned date. As was already been discovered, potential stakeholders tended to be busy and an invitation by a research assistant would most likely go unnoticed and unreacted, the actual sending of the invitations was conducted by the head of the SESP-project. The invitation list consisted of 29 persons, including administrative senior positions, researchers, professors, and IT-related personnel. Actual turnout was seven participants. The workshop was planned as a two-hour session, utilizing the method by Yoo (2018) the researchers were already familiar with. In figure 14 is depicted the setup of the workshop.



Figure 14. Workshop setup.

The method was once again modified. Tokens were replaced by post-it notes as there was no budget available for relevant tokens for a larger group. Participants were exhorted in an opening brief to first write down individually relevant actors, inside of the University of Vaasa or outside of it. Actors could be university organization related, such as teachers, research groups, or academic units, government-related organizational entities such as ministries or parts of the local municipal organization, business, other universities or IT-related. In short, everyone and everything they could think of as related to a Big Data platform.

After each participant had actors they recognized written down, participants were divided randomly into two teams, each with their own area to collaboratively design their vision. First, the participants went through their individually discovered actors one by one and discussed them, removing overlapping actors and merging possible different insights about an actor into one. As more and more actors were identified, they were initially grouped to the whiteboard area provided for each team. After groups had formed, the participants draw lines representing the relationships between the groups and described the relationship with yellow post-it notes. With red post-it notes, they could describe any kind of idea, thought or vision that was not suitable for expressing otherwise. Initial grouping could be re-examined and iterated in any phase. The goal of the workshop was that with this kind of participatory and co-operative method voices not yet heard of in this study would be reached, and greater insights would be gained on how the platform and the related entities were envisioned operating in the context of the university organization.

As the participants were asked to describe the relationships with additional notes, it was also hoped these could provide material where harms or benefits for the entities could be inferred, especially on relationships where the participants had the first-hand experience, but this turned out not to be the case. A possible explanation could be that in this kind of group-based situations, thoughts that can be thought of as relatively personal are difficult to express, especially in the presence of work colleagues. Moreover, the facilitation performed by the researcher could have been more effective in guiding this kind of expression. Lastly, considering the work process during the workshop, the post-it notes proved to be an inadequate change in the method by Yoo (2018), mainly for practical reasons. They did not allow for fluent enough iterative placing of actors, as the tokens did. It is suspected that this had also effects on the results.

5.7.2 Workshop results

The workshop resulted in two different idea maps related to BD, by two teams. The first team had four members and the second one three. Results by the teams are depicted in the appendices eight and nine, respectively. These drawings depicted are drawn based on the photographs taken, textual content is translated to English as directly as possible and care is taken in ensuring transferring all the details to digital form. For example, lines connecting the entities have arrows only where the original drawing had them, depicting a relationship with directional influence and otherwise without directional markings, representing the only relationship. Participants used textual additions on describing the relationships sparingly and additional ideas and thoughts to be described with red post-it notes participants only used once.

Yoo (2018: 4) describes her method revealing in one particular case study not the VSD categories of direct and indirect stakeholders, but rather the *core* and the *peripheral* stakeholders. She observed that the core stakeholders were tended to be placed in the middle and the peripheral towards the edges of the sketches. It is interesting to examine the pictures produced by the two teams with this approach. The first team placed the research in the middle, a group consisting of the performers of the research. That was then connected to related entities, suggesting that this team saw the research as a concept as the core of the platform. With the second team the overview was similar, but instead of research, the core was divided into four different entities, each describing different kinds of research projects and the different goals of each. This difference should not be exaggerated, but it is possible it provides a bit of insight into the multiple aspects of the purpose of the research by the participants.

The first team had a clear distinction of funders, performers of the research, and the audience or the users of the knowledge gained by research while also depicting the impact of the platform in the societal and governmental level. The result of the second team can be depicted as more goal and practice-oriented, as it is more concerned on how different projects that provide the necessary funding for the research are related to the platform and research infrastructure. These depictions were used in the creation of another SESP-project deliverable, Big Data strategy paper, while they also confirmed the research and discovery-oriented purpose of the platform.

5.8 Phase VII: Second conceptual investigation

The second conceptual investigation was performed to examine the results of the interviews on a conceptual level. The values most effectively describing the purpose of the platform are identified.

5.8.1 Value mapping

According to Friedman et al. (2008: 88-89) after identifying the harms and benefits affecting stakeholders, these should be mapped to values. Harms and benefits in this study were recognized with an empirical method by content analysis of stakeholders and discussed previously in section 5.6.3 and 5.6.4. Additionally, direct human values were discovered and they emerged in the cluster *Values* in table 6 in section 5.6.2. Here, these are discussed first and then combined to with harms and benefits analysis to form the final table of relevant stakeholder values in play in the context of the platform. Directly emerged values from the interviews are described in table 11. In the table values are understood by VSD definition, “what is considered important in life”.

Table 11. Values interpreted from Values cluster.

Interpreted Value	Quote
Learning	<p><i>“Why do I want to learn new things? It starts by my own volition” ..“It’s because it’s useful for me in the future in the work or elsewhere, or just out of curiosity”. (I15).</i></p> <p><i>“It is one of the realities of modern life. It’s everywhere. As a researcher or in the industry. The world has become faster and globalized. New technology and digitalization require one to be ready to learn and to find out about things. It’s the requirement for developing yourself”. (I13).</i></p>
Trust	<p><i>“..is better to have high security for the data and the company is actually sure the data is secure and only for academic usage. And that way they are interested in sharing”. (I6)</i></p>
Working together	<p><i>“When you are in the topic in an area of study, that way new insights can arise. Nobody from the sidelines or [a lone] method expert can do it, it arrives as we chew it together. It’s networking and co-operation”. (I2).</i></p>
Challenge	<p><i>“You know being a researcher is challenging work. It’s not the same day each day. You can go to a company and work, and do the same work every day forever. But you know when you re-search, you have a new challenge maybe every day, maybe every week”. (I6).</i></p>

Environmental-ism	<i>“Climate change and the fundamental values should be grabbed if we are talking about the Vaasa energy sector and that business. It is one view of course, but then is this societal view, where there’s a lot of discussion at the moment, and climate change is related to that. Energy chain is related to it, thinking as a layman, very tightly”. (I13).</i>
Equality	<i>“Everyone should have an equal chance of using the platform, so nobody runs it on full power all the time”. (I15).</i> <i>“Every time there’s humans involved.. There’s a chance that someone starts to appropriate the system, for their own uses, blocking others away”. (I2).</i>

Learning was related to two aspects, both of which could be interpreted to be related to egocentric benefits. Firstly, learning provided by the usage of the platform would be able to provide the skills needed in the working life. Secondly, it could be viewed as a necessity to perform well in a modern society where continual change and evolution can be seen as a standard in most areas. Trust was seen as a requirement to gain something interesting and was gained by providing enough security for the platform for the partners to trust in the care of the data. Working together was related to the concept of co-operation emergent in other clusters and interpretation is that it is linked to the recognized know-how gap in the field and in the lack of clear vision of what is actually possible by technological implementations. By acting in co-operation the insights and the knowledge could be merged to gain a more complete understanding. The personal challenge emerged as an opposite to the boring or repeating tasks, it keeps the pursuits of work-life interesting.

There is an additional aspect to personal challenge, based on the background of that particular participant as a researcher. It is rare that routine or boring tasks result in new knowledge, which was described explicitly by the participant to further societal and humanistic goals of improving lives at the level of nations and individuals. Environmentalism is described rather explicitly in the relevant quote and equality is related to the idea that the resource usage of the platform should be controlled and transparent.

From the *Harms* cluster discussed earlier in section 5.6.2, following values related to the system were identified in the classes and categories inside the cluster, and are described in table 12.

Table 12. Values identified in Harms cluster.

Class inside Harms Cluster	Category inside Class(es)	Inter- preted Value(s)	Quote
Control of the Platform	Malevolent usage	Trust	<p>“Then there are these technologies, of another owner and then you can as an outsider to deduct certain things about its efficiency and typical settings”. (I10).</p> <p>“For the most, the industry company is not interested in sharing information to others. It’s very high competition and they work on the novelty”. (I6).</p>
Control of the Platform	Information security	Privacy	<p>“Only threat that comes to mind is that, depending on what is collected and how, but concerning anonymization of information that it does not fail. That it could be traced back to individuals or used malevolently, that this guy answered this way. Privacy is one of the concerns”. (I11).</p>
Control of the Platform	Distribution of System Resources	Equality	<p>“Everyone should have an equal chance of using the platform, so nobody runs it on full power all the time”. (I15).</p>
Usability	The ability of users to use the system effectively, Additional Learning	Efficiency	<p>“Harms and some troubles if the system is complex or unstable, the work kind goes to achieving a simple thing or the system goes down and one has to wait”. (I10).</p> <p>“That is as simple to use as possible for all, so there’s no need to create external [education] systems to enable the usage”. (I9)</p>
Usability	The ability of users to use the system autonomously	Autonomy	<p>“Harm can be if there are not enough user-friendly ways to make use of the data, analyze it and get results out. Do we always require support or some algorithm for it?”. (I12).</p>
Lack of Know-How	Education usage of the platform	Learning	<p>“Harm can be that the system is bad and does not represent the real world, then it does not teach the right things, possibly even the wrong thing, and structures”. (I15).</p>
Lack of Know-How	User adaptation challenges related to operational lack of knowledge	Objectivity, Efficiency, Utility	<p>“Exacerbating, it could be that we can utilize BD in the correct way. That research and queries made would be done [methodically] properly, statistically properly. You know, lies, lies, and statistics”. (I1).</p>
Lack of Know-How	Incorrect Expectations, Resistance	Understanding	<p>“It’s not a black box that you just pour data into, you get new things out [...] I would think the knowledge inside the organization of what the BD is and what it means, does not exist. [...] In the beginning, it may cause bewilderment and resistance”. (I7).</p>
Lack of Know-How	Unknown technological area	Understanding, Rationality	<p>“I see this whole area from the perspective of use being on the unknown ground and that there does not exist a single right way to see it”. (I2).</p>
Material related	Data ownership related issues including ownership of data collected by devices, ethical, contractual and juridical concerns	Informed consent, Privacy	<p>“Then there is the privacy, like for example GDPR, how it prevents information collection and utilization. Can the BD be used if not separately asking each and everyone for permission?”. (I12).</p> <p>“Possibly not harm, but a challenge related to the system. Who has the right to use the data? Some of the information can be public, some not”. (I6).</p> <p>“Electrical grid operating company owns the meter, takes it to the customer but they are not certain if they own data [that meter collects] or is it the customers”. (I10).</p>
Material related	Data collected covering only one research area	Equality	<p>“I have been a bit bothered that the whole of the system has been built for only to be able to simulate various things”. (I12).</p>

From the emergent cluster depicting Potential Benefits for the stakeholders, discussed in chapter 5.6.2 in more detail, the following values presented in Table 13 were interpreted.

Table 13. Values identified in the Potential Benefits cluster.

Class	Category	Inter- preted Value(s)	Quote
Based on Data	Storage of Data	Learning, Knowledge	<i>"[saving and the storing same data for long periods] would be really good, because we could see completely new things appear". (I11).</i>
Based on Data	Sharing of the Data	Co-operation, Sharing	<i>"Then one possible thing would be something that is used by outside entities. That there is something open data that is used by someone, or data that is sold or can be contracts based on us having data available". (I16).</i>
Based on Data	Availability of Specialized Data	Curiosity, Objectivity, Efficiency	<i>"We have two kinds of data. One of them is general data that we can find on the internet, or by experience, we can arrange it. But we need some data that is very special and technical. [...] and that industry shared some of this data, I know we just got it". (I6).</i>
Education	-	Learning	<i>"Would learn the right methods, would be able to use them and to see that they produce proper data". (I1).</i> <i>"Students should have the possibility of learning to use the system, possibly for own projects, courses or for master's thesis". (I5).</i> <i>"As an example for teaching, It would be the biggest usage need [for me]. Could be as an example to use data or how to use Excel with large datasets". (I14).</i>
Facilitator	-	Environmentalism	<i>"How the consumer behavior changes or could change, how the system could support energy transformation. [...] I have not systematically researched, but could we with this kind of data storage and analytics guide and direct actions to facilitate the change". (I12).</i>
Facilitator	-	Efficient Consumption	<i>"If I look it from the point of view of Smart Grids, how the [energy] consumption can be guided or directed with this BD. With it, we can anticipate monthly or weekly changes in consumption or production". (I12).</i>
Facilitator	-	Creation of commercial activity	<i>"As versatile as possible, flexible that it would serve as well as possible the development of future business models, projects analyzing customer behavior, for example as I work myself in the energy area". (I10).</i>
Co-operation	-	Efficiency	<i>"We have the focus on the research and it is largely research instigated this BD platform. But if we think so that it is one central part of this Smart Grid laboratory, that is a part that connects Energy Lab, Engine lab and Smart Grid lab and others". (I10).</i>
Co-operation	-	Sharing of experiences	<i>"It has always been the strength of the university community, supporting services, IT, researcher networks and national networks that we always share openly". (I2).</i>

These previously presented interpreted values from the empirical data are not tied to any universal human values with ethical import, rather, these are things and matters that the participants are interpreted to consider valuable in their life, especially in the context of the platform and their platform related goals.

Additionally, in VSD there should be made a distinction between explicitly supported values and the stakeholder values (Friedman et al. 2008: 82). Previous values discussed are stakeholder values, interpreted from the empirical results of performed interviews. Additional explicit values exist and can be interpreted from the original SESP-project plan by Antila et al. (2016).

The main purpose of the project must be established from the main project document, which is discussed in more detail in section 5.1. This purpose of the project can be condensed to the development of Smart Grid related technologies and concepts in order to facilitate energy transformation to renewable energy sources. It is interpreted as an explicit value of environmental sustainability, as that is the value-based reason for renewable energy production.

5.8.2 Identification and investigation of final values

Values related to the platform, either explicitly or inferred and interpreted from the interviews by content analysis are now identified. As a result, quite many values that participants saw important in their daily lives and related to the context of BD and analysis were revealed. Based on these, human values with ethical import best describing the purpose of the platform were interpreted, condensing the empirically discovered values. These are shown in figure 15.

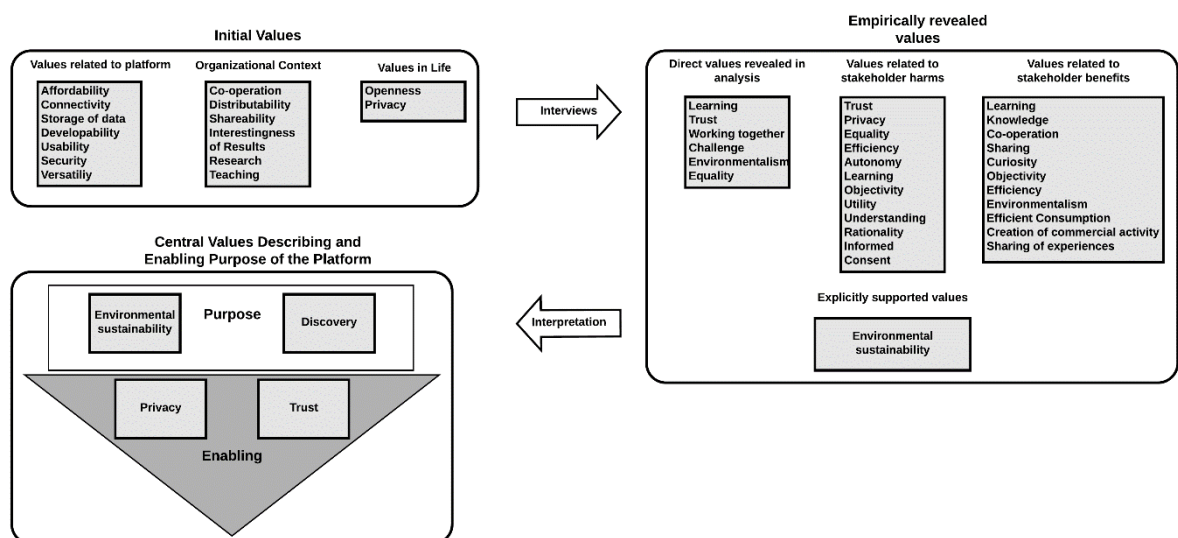


Figure 15. Final value interpretation.

These values can be classified as representing two different types of values. Firstly, as in the case of environmental sustainability and discovery, they are describing the purpose of the Big Data platform. The second type of value is enabling value, which is needed for the purpose, described by the first type of values, to be met. These second type of values are trust and privacy. These four are discussed in detail in the following.

The first value is *environmental sustainability*. It is both explicit value of purpose of the SESP-project and also identified in the analysis of the interviews. Environmental sustainability is described as “sustaining ecosystems such that they meet the needs of the present without compromising the future generations” by Friedman et al. (2008: 91). Energy efficiency, the efficiency of the energy consumption and energy transformation to renewable energy sources facilitated by Smart Grid research is the direct purpose of the SESP-project. Additionally, this was reinforced by the societal perspectives and effects discussed by several participants in the interviews and interpreted either as environmentalism or efficiency of consumption.

The second essential value of the platform is *discovery*. Oxford Dictionary (Oxford 2019) defines discovery as “an act or process of finding somebody/something, or learning about something that was not known about before”. It is a synthesis of several initial values in the context of an organization such as research and teaching, with distributability and shareability of the results facilitating it. Furthermore, it is directly linked to values discovered in the interviews such as learning, knowledge, curiosity, sharing of experiences, rationality, understanding, and objectivity. It describes the purpose of the academic environment the platform is situated in: research, education, and learning. Several participants described reasons they are researchers by the effects of discovery, improving the lives of people and advances in knowledge. An explicit value of the SESP-project, environmental sustainability is only possible with discovery. Discovery, interpreted as a value encompassing the discussed values and objectives, is the second essential value defining the purpose of the platform.

Third value central to the platform, mostly as enabling aspect for the purpose of the platform, is *privacy*. It is understood here as referring “to a claim, an entitlement, or a right of an individual to determine what information about himself or herself can be communicated to others” Friedman et al. (2008: 91). Privacy was identified in the initial values based on reviewing the project documents and by literature review as being a central

concept when considering BD implementations and analytics. In the analysis of the interviews, it was clearly identified as a value by participants. It directly affects the system design via legislation and public concern on the data securing phase, it affects the analysis phase and publication of results. It is central to the operations and development of the platform.

Fourth central value identified was *trust*. Definition of trust by Friedman et al. (2008: 91) is “expectations that exist between people who can experience goodwill, extend goodwill toward others, feel vulnerable, and experience betrayal”. Trust was empirically identified as related to the harms of the stakeholders and as a direct value evident in the interviews. Other values such as transparency and security, accountability, co-operation, and equality either require trust or enhance it. Good security of the platform creates the trust to provide data for the platform, for example. Trust in equal distribution of system resources for different disciplines, researchers and projects are essential to larger scale user adaptation as another example. Trust is in play more as an enabling value for the purpose of the platform than describing it.

5.8.3 Value conflict identification

Value conflicts exist between the identified four central values. These are described in table 14. Mainly these exist in the form of privacy and trust acting in a limiting role, if these were discarded, potentially more gains could be reached in the discovery and to further environmental sustainability. However, even if these can situationally limit the potential gains, neglecting privacy and trust would in the long term do a disservice for the goals of discovery and environmental sustainability by reducing the amount of data and usage efficiency of the platform.

Table 14. Identified value conflicts.

	Environmental Sustainability	Discovery	Privacy	Trust
Environmental Sustainability		Results of discovery processes diminishing environmental sustainability	Data that could potentially enhance environmental sustainability cannot be used or acquired for privacy reasons	Lack of trust for the measures conducted in the system or their end products to further environmental sustainability
Discovery	Results of discovery processes diminishing environmental sustainability		Respecting privacy restricting the source data, education usage, analyzation and publishing	Trust limiting material discovery processes can use in the platform;
Privacy	Data that could potentially enhance environmental sustainability cannot be used or acquired for privacy reasons	Respecting privacy restricting the source data, education usage, analyzation and publishing		Limitations to the platform usage, adaptation and source data availability by the inadequate trust for privacy measures
Trust	Lack of trust for the measures conducted in the system or their end products to further environmental sustainability	Trust limiting material discovery processes can use in the platform;	Limitations to the platform usage, adaptation and source data availability by the inadequate trust for privacy measures	

Value conflicts should not be considered as “either/or”-situations in VSD. Rather, they should be thought of as limitations on the design space. These presented considerations should be integrated into the organizational structure, which is achieved in this study by forming design principles to guide the development of the evolving platform. (Friedman et al. 2008: 91).

5.9 Design principles

Based on the identified essential values, results of the interviews, technological investigations, and presented literature regarding BD, effects of BD in the organizational context, knowledge discovery from data and the software available, the following seven design principles are suggested as guiding lines of evolving BD platform development, not in any order of preference:

1. Identify the purpose of the platform
2. Ensure versatility
3. Creation of and caring for Privacy and Trust
4. Plan for the connectivity
5. Use modularity
6. Serve the users
7. Evaluate and improve, pursue design goals

These principles are discussed further here and their implementation in SESP-project is described in the following chapter six.

Identify the purpose of the platform

This is suggested as of utmost importance. It is based on the hypothesis that the purpose of any artificial object describes and condenses the essence of the object. It is the reason why it exists. The basis of a line of thinking is in how Simon views the artificial, that the artificial things can be described by their functions, adaptations, and goals (Simo 1996: 6). As useful prediction of future is a perilous task, instead of risking an unworthy attempt of it and planning for all the possible paths, much better is to re-represent the design problem which Simon (1996) also advocates: instead of guidelines for all the possible paths, find the core reason why the artifact is in existence and use it as the basis of adaptation in the unknown situation. In SESP-project, this purpose was identified with VSD analysis and the resulting values describing the purpose of the system are environmental sustainability and discovery. Design decisions made in the future should be evaluated if they further or hinder these purposes.

Ensure Versatility

Basis of this principle is largely in the previously presented literature, especially Begoli & Horey (2012), in the situational exploration of stakeholder priorities, requirements and values conducted in the phase V by the interviews and based on the available technology, both by examination of it in the technological investigations and the examined continuously shifting best practices of the practitioners. Essentially, in this principle are synthesized two of three principles presented by Begoli & Horey (2012): *Support a Variety of Analysis Methods* and *One Size Does Not Fit All*. These are discussed in more detail in section 2.1.2. Technological investigations revealed the complexity of component interaction, the different technological requirements of different analytical approaches and the possibilities of potential within a versatile approach. From the best practices examined during the technological investigations, the multitude of configuration approaches became evident and the situational effectiveness of each. To ensure design can be adapted to changes in the outer environment, the inner environment must be versatile enough to allow adaptation. Moreover, the versatility requirement was apparent in the conducted interviews. To be able to serve the various disciplines of the university and thus the previously discussed function of the platform, each discipline with different approaches and requirements in their discovery processes, versatility must be provided.

Creation of and caring for Privacy and Trust

This principle is mainly based on empirical and conceptual investigations and their effects on the technological level. Privacy is understood to be related to the BD phenomenon even in the popular discussion. It was identified as a value in play already in the initial investigations. Furthermore, it was revealed to have direct effects for the data usable in the platform, for example in form of legal restrictions concerning data related to individuals. Anonymization of such data is required. Trust has several aspects. Most critically, security of the system is understood here as mainly a tool to build trust and to ensure privacy. Trust affects the user adaptation, both in using the system and providing data for the system and the storage of the said data within. Both of these aspects require constant upkeep during the evolution of the system, both on the organizational and technological level. Both Trust and Privacy were identified as possibly conflicting with the purpose of the SESP-platform. These conflicts can be alleviated with the development and upkeep of proper control procedures in the platform. It is strongly suspected that in relation to data analytics, there would exist value conflicts with BD platforms build for other purposes, hence making this value conflict generalizable in the field.

Plan for the connectivity

This principle is based on the literature discussed, needs of the stakeholders revealed in the empirical investigation and in identified possible directions the BD paradigm might be heading. Firstly, within is integrated the third principle proposed by Begoli & Horey (2012) *Making Data Accessible*, which is discussed in more detail in section 2.1.2. In the second empirical investigation, it was revealed that many benefits and positive value generated by the platform were directly related to the availability of the data sources. Even if this insight is situational for the SESP-project, the same is true in general for data analytics platforms. Unless all the data is generated within the system, it has to be transferred either in batches or via real-time aggregation. Several existing repositories publicly available data offer accessibility of data via APIs and this amount is expected to be increasing. Additionally, cloud service based virtualization of data processing clusters seems to be a growing trend, and I would expect it to create needs for the ad-hoc clusters to access the data and lead into increased need of transferring the resulting data sets.

Use modularity

Not a new insight in software development or architecture, but the need is based on new grounds. In software architecture, the preference of the modularity of components is based on the efficiencies gained in QA by using modular design and in the additional emphasis on the architectural considerations enabled by usage of modules (Koskimies & Mikkonen 2005). Here, emphasizing modularity is only partially based on these grounds. Ensuring versatility, developability and connectivity require modularity for evolving platform. New components develop and evolve continuously in the Hadoop ecosystem. Additionally, it is strongly suspected based on the technological investigations and literature that as evaluation and real-world testing progress, new needs required by the actual real-world work and the organizational environment arise. Modular design is the only way how to incorporate these anticipated changes.

Serve the users

I would suggest that no socio-technical artifact is able to fulfill the purpose it was built for, unless the tasks performed by the human components are optimized with the understanding of the bounded rationality of the humans, to frame the issue with the term of

Simon (1996). This was evident in the results of the interviews, both qualitative and quantitative. Usability can be thought of as having the responsibility to provide the necessary partitioning of complex wholes into parts or hierarchies that can be effectively internalized by humans. Usability was highly prioritized and in closer analysis turned out to contain several different aspects. Essential aspect discovered was that usability concerns could be interpreted to be resulting from efficiency concerns. If the platform does not provide well-thought of organizational processes for actual usage, the user skill-requirements are not tackled via offering familiar tools or languages for operations, nor the information and guidance in and of operations are not sufficient, it will clearly result in suboptimal user adaptation, if not actual resistance. Resistance hinders the fulfilling of the purpose of the platform.

Evaluate and improve, pursue design goals

The iterative system design methodology is not a new idea by any means. The resulting agile methodologies have been used in computer science and been proven useful and efficient with several of them developed and seem to be continuously evolving into improved versions (Serrador & Pinto 2015). It is not new in IS either, as several DS methods reinstate the necessity of iteration and feedback loops such as DSRM by Peffers et al. (2008) or the solution-based probing by Briggs et al. (2019). Complex invisible and intangible systems are by their nature extremely difficult to understand as a whole, to see and understand all effects and consequences. Therefore, it is no wonder why progress and evaluate loops are predominant in many information technology-related fields. The ever evolving system by default should have that approach. Moreover, the BD analytics as a field with immature software solutions, components, and usage processes reinforces that conclusion. BD platform cannot be a fixed solution, it must be continuously developed, maintained and evaluated to fulfil the purpose of the system. Furthermore, iterative development should have defined development goals, to evaluate the development process. It is entirely possible that these design goals have to be assessed too, in the light of new knowledge. Design goals can be described as a flexible guiding post showing the path to finished system fully serving the essential purpose of it, but it can well turn out that the path does not go the way initially imagined.

6 DEMONSTRATION OF DESIGN PRINCIPLES IN SESP-PROJECT

This chapter is mostly SESP-project related documentation. Additionally, it serves the thesis by providing a demonstration of the developed design principles in practice and their implementation in a real-world project. Evaluation of the system in practice, a necessary step, is left for the possibly following future studies of the platform performance, process development, and continuous evolvement.

6.1 Alternatives and arguments for selections

As operator partner left the project there were not many options left for prototype development. Two alternative routes were left for the development of the prototype: cloud-based solution or on-premises prototype running on owned hardware, possible sources of said hardware being unknown at the time. Both approaches were limited by the budget available for purchases, which at the time was nonexistent.

Cloud-based platforms were briefly investigated by a combination of free usage credits provided by platform operators and personal credits accrued by the researcher from other sources. Amazon AWS, Google Cloud and Microsoft Azure environment were explored by creating and running Hadoop cluster within. Additionally, Hortonworks provided packages were tried to deploy where possible in order to provide a similar component palette to the first prototype.

The primary benefit of the cloud-based approach is cost efficiency. All the examined platforms were based on the idea of offering cost savings compared to the acquisition of on-premises hardware. It is evident and general knowledge that the capacity of a purchased computer cluster is never *exactly* right. There is overhead in requirements to allow unforeseen changes, operational developments, and changes. It is a common presumption that capacity is always underutilized or the capacity limits the necessary utilization, as the capacity is increased in different sizes of fragments, not smoothly. That combined with the different levels of complexity in procurement processes might lead to the situations where overcapacity is preferred. In contrast, in cloud-based solutions capacity can be always exactly right. Costs are generated by the exact usage. If more capacity is needed, it can be bought as long as it is required, in the solutions of all three platform providers.

This scalability is available for both the computation and for the storage services. Additional computational needs can be fulfilled by using more powerful virtual machines inside the cluster, using more of them or, if the additional computational power is required by different use process, by spinning up a whole additional cluster. Cost-efficiency is provided by the billing being based on the uptime of the cluster multiplied by the hourly rate of the used virtual machines. Obviously, virtual machines offering more calculation performance are more costly. Once the calculations are done, the cluster used is powered off and if required, removed entirely. Creating a new cluster exactly when the need arises is rapid enough with all three major providers for this process to be of use.

Data storage in the cloud follows similar principles, costs are based on the amount of data stored and the number of operations performed against it. Storing larger amounts of data that is processed actively costs more than storing smaller amounts of inactive data. Created clusters have differently implemented ways to access the stored data and to store the results of processing in the three services. Of the three briefly investigated services, Microsoft Azure offering working HDInsight component was identified as the most suitable for deeper investigation. Azure Datalake Storage generation 2 seemed to offer necessary Hadoop compatible file system endpoint integrated into Azure Blob Storage. Generation 2 storage was in the preview phase during the initial investigation.

Monthly costs would then consist of a combination of analytic and data storage costs. These are presented in the following table 15. Cost of approximately \$5500 for the usage of the following cluster is an estimate, more exact costs would have been revealed by active usage and testing. Costs presented do not represent costs involved in fully working BD platform in operational use. They represent costs involved in fielding a proper prototype to investigate real-world usage in a cloud environment.

Cloud platform as a basis for the BD platform would need further examination and evaluation in real-world usage. I would estimate such evaluation would take at least months and most cost-effective way to pursue such an investigation would be starting a consultation process with a suitable larger organization with some level of partner status in the project.

Table 15. Cost evaluation of cloud-based prototype 2.

Monthly Cost	Item
\$566.40	2 x Head Node (D5: 16 cores, 56GB RAM, 800 GB Temporary Storage, \$1.18/hr)
\$3588.48	10 x Worker Node (D14: 16 cores, 112GB RAM, 800 GB Temporary storage, \$1.50/hr)
\$1,342.60	Average storage of 60 TB per month
\$5	1,000,000 API operations
\$5	1,000,000 List and Create Container Operations
\$0.4	1,000,000 Read Operations
\$0.4	1,000,000 Other Operations
\$5504,28	Total Monthly Estimate

Analysis of the cloud-based solution against the heuristics of the design principles has to be considered extremely initial and an estimation, as the investigation was severely limited by the usage credits available for the researcher. However, initial conclusion was that privacy and trust, being universal issues related to data-oriented research could be provided – security and control of user access were fine detailed but issues could arise in the areas of versatility, modularity, and connectivity. Platform locked development would mean limitations in those areas in addition to the capability of serving the users. Development of a cloud-based BD platform addressing all those concerns should be considered only with enough resources to develop it in co-operation of an outside organization having sufficient development resources and knowledge of the operations environment.

Building and developing platform prototype on on-premises hardware seemed not to be possible, but as discussed in more detail in section 5.4 an opportunity arose to pursue that path of investigations further. As the cloud platforms were already initially evaluated and their constrictions and know-how related challenges were known, they could be compared to the understanding of the environment gained in building the first prototype. It was known that on-premises solution would enable building a much more versatile and mod-

ular stack of different components and allow for a much larger catalog of actions to address the serving of users, as almost everything could be tried and done. Connectivity related issues could be more difficult to solve, however. Moreover, and perhaps more significantly, starting of building the prototype would not have to wait for approval of funds and could proceed once the access was granted. Even though the process had a resource related unanticipated waiting periods due to the busy schedule of the IT-management and teething problems with the environment provided, which should be considered normal when trying new things, this path allowed the development process continue.

It is expected the on-premises solution to provide more technical know-how of how to provide for and address the issues outlined in the design principles, with the freedom provided by the on-premises solution allowing far greater opportunities to investigate different components, solutions, and modules. It is presumed that the BD platform prototype will be able to be actively used in versatile research, especially as the additional funds are used in the expansion of the hardware instead of monthly payments to the cloud platform provider. However, if the continually developed and evolving platform in-house platform can provide a solution to the scale of objectives outlined in SESP-project is not certain. It is entirely possible that at some point of the lifetime of the platform the best course of action would be searching for a larger outside entity to provide for the scalability required.

6.2 Technical architecture documentation

The prototype of the platform is documented in this section as it is currently and a higher level picture of the immediately following development plans is provided. The prototype is documented on five levels: the overall view of the architecture, the VM environment level, OS level of the nodes, Hadoop environment level and the external components level. Network level documentation of the prototype as it exists at the writing of the thesis is provided in figure 8 in section 5.4.2.

In figure 16 is presented the architectural view of the prototype of the platform describing the usage process at this stage. Data platform is considered as an entity comprising of HDP core with the necessary components configured. It is considered likely that the platform will include various external components not part of the HDP package. Most likely the first additions will be related to establishing another custom UI for allowing easier access to the results the researchers want to share. Initial plans regarding this are based

on utilization of well-known light-weight open source stacks, deployed on additional VMs in the environment. Sqoop is planned as an external connector, but further investigation is required in the deployment and development phase of the additions. These plans are related to the principles of connectivity, modularity, serving the users, and continuous evaluation and improvement.

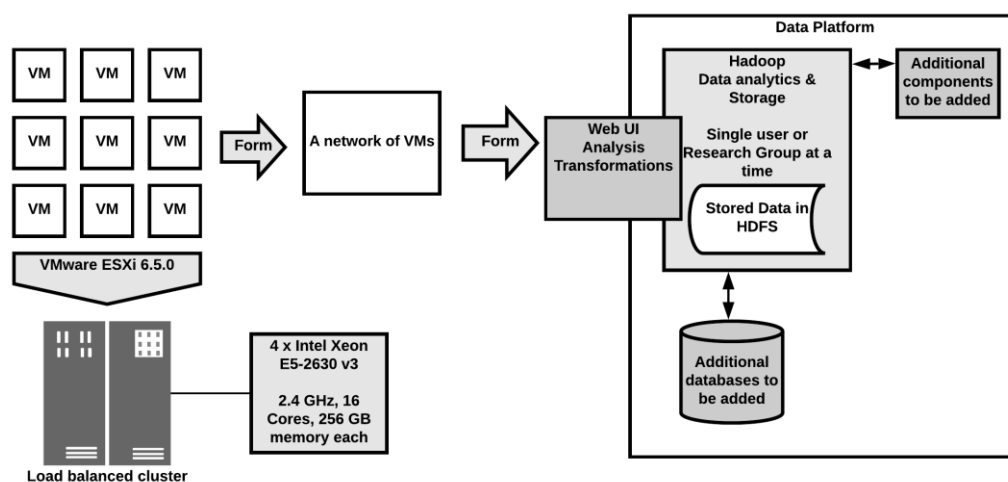


Figure 16. Data platform overall architecture.

Environment the platform prototype is situated in consists of VM virtualized environment provided by VMware ESXi 6.5.0, a type 1 hypervisor software, which directly controls the hardware available and manages resources for the virtual machines deployed (VMware 2018). It has available a cluster of four Intel Xeon E5-2630 v3 CPUs, operating at 2.40 GHz each with 16 logical processors and 256 GB of memory. In the VMware environment saving capability was provided by six usable data storages. Four of the data storages had 4 TB available capacity and two provided 1.5 TB each.

Internal storage of the two master nodes was provided with a single 750 GB each, while the internal system files and capacity used for HDFS in the six slave nodes was divided to three 500 GB hard drives and a single 300 GB drive which provided for the discovered need for additional cache and temporary file space. Even in this virtualization system level, the principle of modularity was followed. Division of the used space into multiple

virtual hard drives allows for versatility in the unknown future – it allows adaptation to the changes on both the virtualization environment level and on the system design level.

As described in section 5.4, the design was based on a cluster of ten nodes originally. After testing and evaluation, it was discovered that the overhead from the number of virtual machines was affecting the HDFS-capability of the prototype negatively at this point. To improve efficiency, the number of virtual machines was reduced. Additionally, as each virtual machine was based on partitioning scheme where the OS, cluster configuration and the HDFS reserved space resided on the same virtual hard drive, it directly violated the principles regarding modularity and versatility on the OS level.

As a result of these considerations, the final operating system level partitioning was created as described in table 16 for two master nodes. It roughly follows industry best practices regarding Hadoop-cluster partitioning in sizes and takes into account the recommendations concerning file systems, but is customized for the prototype environment and resources.

Table 16. Master nodes partitioning table.

Hard drive partition	Size	File System	Mount point
/dev/sda1	1014 MB	XFS	/boot
/dev/sda2	197 GB	EXT3	/tmp
/dev/sda3	171 GB	EXT3	/var
/dev/sda6	99 GB	EXT3	/home
/dev/sda7	99 GB	EXT3	/usr/hdp
/dev/sda8	50 GB	XFS	/

Slave node partitioning follows the same principles, but additionally, there exist partitions designed for HDFS storage and the discovered need for additional temporary space for Yarn cache. It allows straightforward expansion of HDFS capacity by mounting additional hard drives and provides enough space for both the OS and HDP upgrades and updates. This is presented in table 17.

Table 17. Drive partitioning of slave nodes.

Hard drive partition	Size	File System	Mount point
/dev/sda1	1014 MB	XFS	/boot
/dev/sda2	50 GB	XFS	/home
/dev/sda5	99 GB	EXT3	/usr/hdp
/dev/sda6	99 GB	EXT3	/var
/dev/sda7	82 GB	EXT3	/tmp
/dev/sdb1	493 GB	EXT3	/grid/1
/dev/sdc1	493 GB	EXT3	/grid/2
/dev/sdd1	296 GB	EXT3	/localyarn

Evaluation of the suitable operating systems was already concluded in the SESP-project and the decision was made to use 64-bit CentOS 7, which was utilized in the virtual machines of the cluster with a kernel version of 3.10.0-862.14.4.el7. Upon that base, HDP version of 3.0.1.0 was installed.

HDP stack initially installed consisted of Hadoop-components and versions presented in appendix 6. More in detail component distribution of the cluster at this stage is presented in appendix 7. It is expected the component distribution will evolve during the development as new possible limiting factors in the setup are identified. However, the stress tests concluded point to the stability of the current configuration, although it is possible all potential optimizations have not yet been discovered. Moreover, the current component palette installed is consciously too large. It is so to ensure the versatility of the platform as it is being developed and it is expected to narrow down as more experience is gathered in practice and the more outside modules are installed. Configuration details are omitted at this stage for brevity but will be presented in more detail at the final report.

The first iteration of prototype 3 consisted of 10 VMs, but after evaluation of performance and the ratio of resources consumed by the infrastructure versus the resources available for the end users, trying to conform to the principle of user-based thinking. The final resource related configuration of prototype 3, consisting of two master nodes and 6 slave nodes is depicted in table 18.

Table 18. Resource distribution of prototype 2.

Item	Master Nodes	Slave Nodes
vCPU	6	10
Memory	64 GB	58 GB
HDD 1:	750 GB System	500 GB System
HDD 2:	--	500 GB HDFS
HDD 3:	-	500 GB HDFS
HDD 4:	-	300 GB System

6.3 Data-oriented architecture documentation

The data-oriented architecture of the prototype is relatively straightforward at this phase of the evolution of the platform. As there currently exists severe limitations regarding storage capabilities of the platform, and the level of control required by the principles of privacy and trust is not yet live in the platform, there is no existing capacity nor immediate plans of long term storage of interesting data. Data-centric current architecture and process model of the platform is depicted in figure 17.

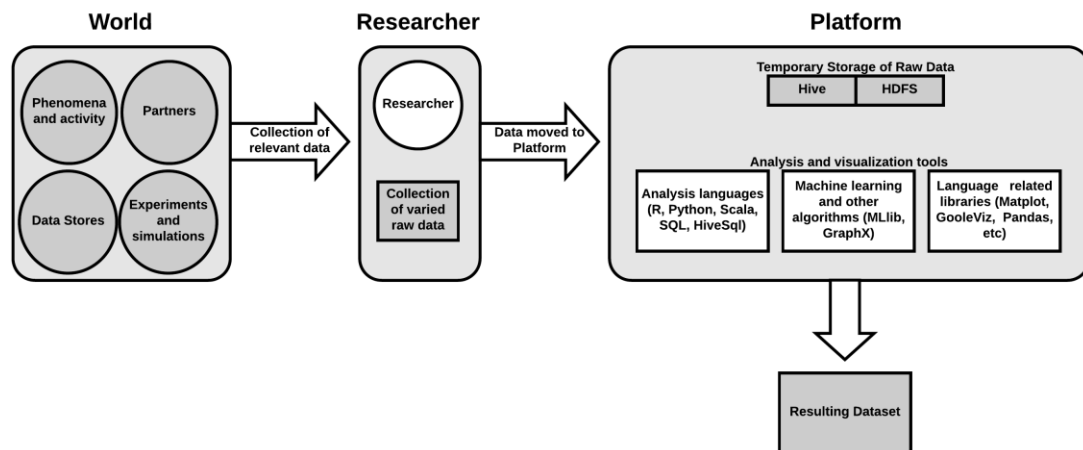


Figure 17. Data-centric overview of the platform.

Start of the process is that the researcher identifies the relevant data sources for their research in the outside world. These can be data related to natural phenomena, activities of various entities including sources such as activity in social media and financial or or-

ganizational activities of businesses or institutions, data provided by partners, data existing in various publicly accessible databases and in other databases available for the researcher, or data source can be results of experiments, simulations or interviews. Data can be in various forms, including structured, semi-structured and unstructured. After the researcher has collected the raw data they are interested in, it is transferred to the platform

Transferring the data to the platform can occur in multiple ways. If suitable, Network File System may be applied or more physical methods can be used. It is an assisted process. A data staging area, where interested researchers may gather their data will be evaluated in the future, after the updates to the system storage have been implemented. It should provide a possible solution to the problem of storage in the data collection phase, more in line with the design principle of serving users, especially from the usability viewpoint. Furthermore, it is in line with the principle of planning for the connectivity. As it is, the raw-data collection is done to storage space provided by the researcher.

After transferring the data is complete, data will be initially located in the HDFS storage. Depending on the type of the data and the specific research goals and planned analyzation methods of the researcher, none, some, or all of the data can also be loaded to Hive. Data loaded to Hive is accessed via HiveContext in Zeppelin. Once raw data is imported, the analyzation phase can begin. Necessary transformations are performed to the data, addressing either analyzation prerequisites or data integrity issues. After these are addressed, the researcher may use the various libraries available, including machine learning algorithms, to analyze the data and build various models. Eventually, the researcher has their data in either final form for the research or in a suitably processed form for them to continue the research process outside of the platform.

The resulting dataset from the process will be extracted from the platform for the researcher. Raw data and various stages of transformations will be removed from the system in this phase to address the storage limitation issues. It is intended to further develop this phase according to the principle concerning connectivity. The first stage would be establishing of platform module allowing storing of ready datasets in binary packed form for distribution, for the audience the researcher chooses. As the development of the platform continues, the principles uncovered will continue to be observed regarding the data-oriented architecture.

6.4 Platform future

This section describes the evolution of the platform, examines the threats recognized, the benefits the platform could provide, and included are most of the concrete suggestions and ideas the participants in the interviews described. This section is mostly based on the six other emergent clusters (three clusters were analyzed in more detail in section 5.6) discovered during the analyzation of the interviews, combined with the directions advocated by the identified and suggested design principles.

6.4.1 Platform evolution and the lifecycle

In general, the view of how the platform development should proceed was relatively unified. The complexity of the final platform was mostly understood, and the knowledge gap related to the development, planning, operating and utilizing it was generally recognized. As a means to reach that goal, only one strategy was evident in the answers. Approach consisting of building a prototype which is then continuously developed, maintained, evaluated and most importantly, used.

Continuous evolution and development based on an evaluation of the existing prototype in real-world settings would offer significant benefits. It would firstly enable proper adaptation to unforeseen changes, developments, and opportunities. This is especially important as the technology itself as a whole, Hadoop components, algorithm libraries, tools and methodologies in discovering knowledge from the data are not mature. They are still in a process of rapid change, even when considered in the context of fast-paced evolution of the IT field in general. It is certain that new possibilities will emerge based on this process alone, but it is uncertain what kind of possibilities these will be.

Furthermore, there can be organizational changes that require new adaptations or new processes. These can be related to the surrounding organization of the university, new research projects, new laboratories, new living labs as data sources or new initiatives. To take advantage of these possible developments and perhaps facilitate them, agility offered by the iterative and cyclical approach, as suggested by the principles, would be essential.

Secondly, it would make the development of the platform to be based on real experiences and real-world practices. This would help bridge the previously discussed know-how deficiency apparent in the goals of the whole SESP-project, in development and in operating the platform. Needs and goals would be based on reality, targeting requirements arisen from the daily use and evaluation of the results of usage and the processes used. There exist numerous examples of BD-related projects that have failed, and it would seem intuitive that an approach based smaller iterations would offer possibilities to reduce these risks.

Thirdly, this approach would foster knowledge related to the field within the university in a tangible way. It is a field containing skills that are highly sought after by the students, researchers, and professionals. An existing, used and continuously developed platform would have a concrete potential for various educational use. Furthermore, by starting relatively slowly, having the complete focus on a single use case at a time, would ensure that all feedback and discovered inadequacies would be noticed and could be corrected. This would result in the user experience being positive by addressing these shortcomings in a co-operative way. The fostering of the related skills in the university would happen in both sides of the equation – both on the side of the users and on the development and maintenance side. Especially important is the gradual propagation of the positive user experiences and the related knowledge of operating the platform to have the university, and the related partner organizations, achieve the potentially high rewards of data-intensive science.

Fourthly, it would be an approach with relatively small financial risks. Instead of one costly and most likely long project with high chances of failure for reasons previously discussed, the baseline of funding could be relatively low if the more iterative approach would be used with evaluation checkpoints and proper design road mapping and milestones. Especially if the early prototype and the following initial development cycles would be built upon the on-premises hardware already existing, as done in SESP, the funding for investments could be also done based on requirements born from experience in the operations of the platform. It would also offer a chance for continuous examination of results achieved with the funding and opportunity to tie funding to certain milestones or design goals, as the principles suggest.

The iterative approach to systems development is not without risks. It requires clearly identified and articulated long term goals. Each development cycle should proceed towards these, accounting for the evaluations of the previous cycle results. It also should have metrics to measure if the development has failed. In the interviews, one particular failure condition rose. The participant saw the system failing when it no longer matched the practices in the industry. Most other risks in this approach are related to the actual development itself, obtaining the relevant know-how, and to the uncertainty of the direction of the fast technological development in the field and in IT in general.

6.4.2 Data sources

Data sources that could be used in the future with the platform were suggested by seven participants, some of them identifying several of them. Some of them were directly in-line with possible data sources already identified in the SESP-project and reinforcing the need for them.

These include data from living lab sites, such as Sundom Smart Grid and sensors within such sites, providing data related to technical aspects of the energy grid. Related is the data provided by companies in the field. Here noteworthy is to notice again the importance of the proprietary data provided by these companies. As one participant put it, availability of such special data would open up possibilities of research unavailable elsewhere, offer real efficiency gains compared to resorting to trial and error to get the required data. Sourcing data from simulations and experiments were also recognized in SESP-project and reinforced by the participants. To complete the technical aspects of energy-related data, real-time data related to the consumption and production of energy, and the data related to the markets of energy were identified. Consumption data especially had two different sides to it, the technical side of seeing it as creating the production need, and the customer-oriented view based on the energy consumption of households as generating behavioral data to be analyzed.

Additional sources of data as mentioned by the participants that were not new, were Electrical Grid Disturbance library, various existing data banks, and collection of data from social media. New insights regarding the data available were the potential usages of the data generated by the technical infrastructure of the Vaasa University. Examples of such are server logs, firewall logs, and routers. It was also recognized that the administrative

data related to the students and their studies could be useful in developing both the teaching and study guidance, and it could potentially offer administrative benefits as well.

Systems in use – Lukkari, WebOodi, and Moodle – provide a lot of interesting data related to the study results and usage patterns of students. Some examples mentioned that potentially automatic study guidance could be improved and developed by improved usage of these data sources. Additionally, student behavior data in those systems, especially in Moodle, could potentially be analyzed and compared with the goal of developing course layouts and gain metrics on self-studying of material done by the students. There exist multiple interesting questions to pursue based on that kind of material, such as how time spent in the course area affects their scores, what is good teaching material, how the quality of the material and layouts affect student usage of the materials and what kind of results all these effects have on their final scores. Administratively, data provided by these sources could perhaps be analyzed and results gained of work ratios to ECTS credits, for example.

All in all, there was a wide range of data source ideas to be integrated into the platform. However, there are three problems in this area. Firstly, at the prototype stage, the storage capacity available creates limits of what can be stored for use. It would be essential that the platform can be used in the analytical capacity, to offer the capability to analyze larger scale data. As the storage resources are limited, I would suggest it to be prudent to prioritize securing of large enough working space in the platform to enable the analyzation of large scale datasets and only secondarily prioritize the storing of data in the initial phase. Additionally, if possible, dedicating some of the storage capacity to serve as a staging area, where the researchers could collect their material preceding the import phase to the system should be investigated.

Secondarily, there exists the question of batch data vs the real-time data what many of these data sources and utilizations require. This is a distinction and an aspect with a great many real-world implications that is absent from the SESP-project plan. Batch data analytics is greatly more straightforward to implement and undoubtedly the most suitable place to start building the prototype. With real-time data analysis the required surrounding physical infrastructure increases in complexity as does the according to software architecture. With current limitations, real-time data analysis and aggregation is out of scope and has to be investigated and developed in the possibly following iterations of the prototype, and the relevant resource and knowledge gaps need to be addressed.

Thirdly, it has to be questioned the value of duplicating open data sources. It is impossible for the platform to provide integrations of all possible data sources that could be of interest for a wide range of researchers. Perhaps better and a more sensible process would consist of the researchers in their data gathering phase utilizing these data sources, several of which offer APIs for retrieval of information, collect the data they are interested in and then in import phase move all their raw data to the platform for analyzation.

With these points discussed, there exist some sources of data that are only available in the context of work and projects done in the Vaasa area, a prime example being the Sundom Living Lab, and the availability of which could provide value for the platform and for the use cases. As the developing storage space allows, investigation on how to most effectively incorporate these as a batch copy in the platform and what kind of experiences result, will be conducted.

6.4.3 Challenges

There are exists several challenges identified by the participants of the interviews, based on the design principles and the literature knowledge encapsulated within them, and as a result of the technological investigations. Most crucial of these are discussed in this section.

Versatility required of the platform will translate into real architecture as a multitude of implemented Hadoop components and outside modules. Keeping these up-to-date, communications between them working, and especially the initial development of the architecture can provide significant needs of technological and theoretical know-how. This can be addressed partly by securing enough resources for the development, outsourcing the development to outside partners or by adopting an iterative approach to the development, as suggested by the principles.

Control and security of the system, especially as constructed as a stack of various technologies, with various ways of inter-component communication, will provide deep challenges to deliver a controlled and truly information secure platform. As the platform will include a multitude of separate technologies, each with their own potential vulnerabilities, the soundness of the security of the overall architecture must be emphasized. It is suggested to review experiences of published data analysis platforms to ensure an approach

in line with the best practices of the industry. Additionally, in the design of the platform, the already adopted approach of offering a minimum amount of user interaction points until the security of the whole can be evaluated, should work towards minimizing these risks.

Legal, contractual and ethical concerns related to the material and data possible stored permanently into the system in the future, especially once the system reaches a phase of multi-tenancy are vast. These concerns must be solved by the researchers and related projects before they either permanently store or share their data. One way to make this process smoother for the users of the platform would be creating a suitable legal template covering the usage of the data, the restrictions and describing the lifecycle of the data in question. The aspect of information management from the legal point of view and keeping this tightly integrated with the physical storage and allowed views of the data requires additional research and development. Furthermore, the platform cannot respond to GDPR requests. It is a requirement for the data entered to the platform for storage or for sharing, that GDPR does not cover it. Anonymization is required before entering data to the platform.

Framing the BD system as a “Platform of platforms”, framing established in SESP-project documents can be only described as a long term goal. The development of a platform of such sophistication in timescales available received earlier in the project quite negative response from an industry professional, who also suggested a more realistic approach of building towards that goal by iterative development during a longer time frame. Additionally, the reasons the operator partner withdrew from the project must be considered. Is the design goal sound and does it really facilitate a business case?

I would suggest amongst the largest challenges is the clear definition of the design goals, as suggested by the principles. What is the platform future and how the purpose of the platform will be fulfilled? These design goals must be defined in order to enable evaluation and planning the development, as is suggested, goals of SESP-project might not be the way to extract the maximal value from the platform for the stakeholders, as it might have other utility as defined in the purpose of the platform.

6.4.4 Design goals

Previously in the thesis discussion related to the benefits and goals of the platform have been provided. Purpose of the platform has been identified and defined as facilitation of environmental sustainability via discovery. However, more concrete design goals have to be eventually provided and declared on how to actually get there.

These design goal needs are multifaceted and tied to resources. A design goal is, for example, the consideration should the development progress by a larger outside operator, partner or in-house, or should the process be mixed. There were many concerns regarding the platform by the interview participants. Answers for these concerns should be developed via design goals as enough technical experience is gained via using the platform in practice and plans for the future of the platform are consolidated.

Design goals should be developed to address the following concerns risen in the interviews:

- Will the platform be productized in some way, if so, how?
- Will this development process result in a platform that is cutting edge, or are major steps performed elsewhere?
- Will it be more affordable, more user-friendly or more affordable compared to other solutions?
- What is the value provided by the platform for the related consumption of resources?

6.4.5 Possible practical steps forward

Analyzation of the interviews, workshop results, and thoughts based on the implementation of the design principles in practice, suggest several practical steps that could be implemented in the possible future development of the platform.

There exist a few potential concrete actions already mentioned. To address the legal and contractual issues considering data during the data lifecycle, a legal template should be constructed to be used by all research projects utilizing the platform, even if the data in question is gathered and owned by the researcher themselves. On the technical side, three

actions should be tested – a building of a data gathering area to assist researchers in gathering of the raw material prior to the actual usage of the platform, an outside module focused on testing the sharing of results in practice and lastly, additional expansion investigation of hardware base to allow addressing the need for additional computational power, infrastructure required for testing real-time data analytics and aggregation, and to create a hardware level solution for system backups. These latter hardware related investigations are mostly restricted by the uncertain nature of financial resources. A rough estimation of the magnitude of these investments by IT is 10 000 euros for the additional computational capacity of 18-20 cores, 768 GB memory and the licenses required. The provision of secure back-up of the platform and data, would cost approximately 35 000 euros consisting of hardware and software costs.

Analyzation results regarding the possible practical steps in the future for the platform were gathered in the emergent cluster “Evolution, Strategy, Development” as presented in section 5.6.2. Co-operation could be named as the major theme related to these views and ideas. This theme was proposed and even exhorted as a concrete way to proceed especially by one participant, with two other participants expressing thoughts in similar lines.

In this theme of co-operation concrete steps to take can be interpreted and it is related to the value the platform can provide by facilitating low-level co-operation inside the university community and the local area. This leads to the following three proposed steps. By providing access and ensuring the platform is tested with research conducted by non-technical disciplines of the university, it should cultivate lower level information exchange in the university about the platform and the analysis possibilities, potentially resulting in unseen new research goals. Secondly, co-operation between the various educational institutions, related to the use of the platform, should be resourced and instigated. Similar advantages to cross-discipline co-operation could potentially be reached. Thirdly, once the development of the platform has progressed through first few test cases, the platform should be included as an analysis tool in partner related project, to evaluate and to demonstrate the capabilities of the platform, and to discover possible new partner related synergies.

The theme of co-operation has other aspects related more closely to the idea of marketing the platform and the university. Spreading knowledge about the platform and the capa-

bilities it provides in the university, to develop a user base and to foster further understanding of the field was envisioned to be conducted by the following concrete steps. Firstly, knowledge should be disseminated inside the university. Teaching digigroup, various theme days, university communications department, academic unit meetings and the proposed digimentors of research, which would correspond to the teaching digimentors on the education side, should be approached and utilized to spread the knowledge about the platform and the possibilities it provides, on a very concrete level. One such way could be discussing the results and cases of research performed in test cases during the development. Tutkimuksen ja opetuksen tuen päivä 2019, a seminar consisting of discussion related to the different technological ways and methods to support research and teaching in the university context, was identified as a good opportunity to further these goals.

Secondly, the same goals should be pursued outside of the university. Actions could be directed towards establishing a loose network of professionals interested in and participating in related research and operations of such systems. Analyzation of large scale data sets, problematics related to real-time ingestion and aggregation, for example, are issues that are confronted by other universities nationally. There would exist numerous benefits if these experiences could be shared or some investments could be done together. If the University of Vaasa would be the one instigating the establishment of such a network, there could be some possible gains. A possible way to start this process was described by a participant as organizing a national seminar or another suitable small event to gather interested actors in Vaasa.

7 DISCUSSION

The contributions of the research conducted in this thesis are the suggested design principles and the example of utilizing both VSD and DSR to explore ways to integrate values on design research.

Design principles discovered and formulated an answer to the original research question of “*What kind of design principles represent the value conscious best practices of a Big Data platform?*”. None of the principles can be described as new or never suggested in the field of IT, considered singly. Value of the principles stems from their grounding in both the empirical experiences and experiments, technological experiments especially integrating practitioner knowledge available, and to their roots in literature presented and discussed, concerning the effects and nature of Big Data, the relevant processes of acquiring knowledge from data, and the peer-reviewed publication of earlier principles. As stated, a single one of these principles considered by itself is hardly a discovery, but it is proposed that adhering to and following the proposed *set* of design principles will lead to the creation of Big Data analytical platform, which is well equipped to serve the identified purpose.

Further, the principles presented embody and integrate the values identified with the presented qualitative and quantitative analysis of representative stakeholder interviews. The stakeholder groups were identified with empirical methods with the application of a new proposed method. Initial related values identified were the basis of the design of these interviews and the analyzation of these interviews provided a more empirical approach to examine benefits and harms related to the stakeholders than is the norm in VSD.

As a second contribution of the research, scientific design in the framework of DSRM was pursued. To gain a value conscious result, to serve the development of the platform for the long term, VSD methodology was combined with the DSRM. It also allowed the examination of the purpose of the platform in a more profound way, to gain multi-voiced input of the possible harms, benefits, and possibly overlooked potential related to the usage and to the existence of the platform.

As a result of the iterative application of the various investigations of the VSD design principles were formed and suggested. They are an IS artifact, both in the sense of consisting of the descriptive knowledge, the “how” of DSR, as Iivari (2007: 46) describes:

“how things could be and how to achieve the specified ends in an efficient manner”, and in the sense Lee, Thomas, and Baskerville (2015) describe – consisting of more than just the technological artifact. Demonstrative usage of these principles in SESP-project is provided. Further evaluation requires additional research.

7.1 Related research

This thesis can be seen as being related to several research streams. It can be discussed in the context of value sensitive design, design science, requirement engineering, and Big Data analytics and storage implementations.

The research approach in the thesis is congruent in how Manders-Huits (2012: 275) views the essence of VSD - that is to both observe and identify the values related to the BD platform and to identify the values considered important by the target group. Moreover, part of the VSD critique presented by Manders-Huits (2012: 277) is concerned with the stakeholder identification. In the thesis, this critique is addressed by employing the stakeholder tokens method proposed by Yoo (2018) in a modified form, once by the designers of the system and the second time in a workshop environment.

Mander-Huits (2012: 278) also voices concerns on the empirical methods involved in VSD investigations, particularly on what do the stakeholders mean when discussing particular value. In the interviews performed, this has been tried to address by designing a whole theme on the interviews, where participants themselves describe what they mean with the values they prioritize.

There exist a point of improvement for VSD that is discussed by Borning and Muller (2012), that has been tried to take in consideration in structuring the research process and the research presentation - the strengthening the voice of the participants in writing about the VSD investigation (Borning & Muller 2012: 1129). This has been tried to achieve by the provision of direct quotes by the participants and secondly emphasizing and clearly describing when values or meanings have been interpreted or inferred from these.

An example of VSD research combining DSR methodologies is research conducted by Dadgar & Joshi (2015), where they used the ISDT methodology proposed by Walls et al.

(1992). They employed VSD investigation to provide the kernel theory, meta-requirements and the meta-design for the subsequent ISDT research. I would propose that the approach undertaken in this thesis could be more sound. Firstly, the choice of ISDT is perhaps not optimal for reasons discussed by Gregor & Hevner (2007: 319-323) and the DSRM providing a possibly better framework for conducting the DS research. Furthermore, it does not sufficiently account for the iterative nature of both DSR and VSD. Structure followed in this thesis has a tight inner loop of VSD iteration, with various investigations following each other based on the findings of the preceding ones. Then it should be followed by DSRM iterations of demonstration and evaluation, which could trigger another DSRM iteration, again employing iterative investigations of VSD in the inner loop. Unfortunately, this work cannot be completed in a thesis timeframe and scope, leaving this for possibly following future studies.

DSRM proposed by Peffers et al. (2008) has seen relatively wide use, but to the best knowledge of the author, this is the first time it has been combined with VSD. In DSR there exist the duality of goals, to improve and create theoretical knowledge - the truth as discussed by Järvinen, and to improve and create socio-technical artifacts beyond the immediate research context – the utility as discussed by Järvinen (Briggs, Böhmman, Schwabe and Tuunanen 2019: 5725; Järvinen 2017). In VSD, the goal can be described as identifying various value considerations and integrating them to the socio-technological artifact consciously, instead of them affecting the design and stakeholders unconsciously, without any distinctly laid out prioritization.

I would argue that the combination of VSD and DSRM serves both the purpose of DSR and VSD as tested in the thesis. The validity of reaching the goals of VSD is not in question. Relevant values were identified, examined and integrated into the organizational context via the design principles to be followed in the project. Additional concerns, opportunities and usage cases were discovered with the multi-voiced approach. However, what needs further evaluation is reaching the goals of DSR. If the suggested design principles are generalizable, and the “truth” in the sense as discussed by Järvinen (2017), then it could be claimed a contribution to theoretical knowledge has occurred. The claim of achieving the goal of utility is on firmer ground but not without dispute. The principles have been implemented in the SESP-project, as discussed in chapter 6, but as development is still on-going, further evaluation and demonstration in actual usage would be re-

quired to address the DSR evaluation needs with prototype type of demonstration, as discussed by Sonnenberg & vom Brocke (2012: 11). A weak claim could be made that demonstration has occurred by assertion (Sonnenberg & vom Brocke 2012: 9).

Requirement engineering approaches the original problem space inspiring the research conducted in this thesis from an alternative direction. Laplante (2014) provides a solid overview of the methods. It is based on meticulous planning, usage of different suitable methods such as KAOS (see for example Pommerantz et al. 2012) and more based on the idea that perfect design is possible and it exists already, it just needs to be excavated from the minds of people by systematical employment of well thought-out and well-devised methods. This approach is not without merits. As a passenger, perhaps most of us are happier boarding an airplane or train designed with traditional engineering methods. In the information system development, iterative approach has more merits, as the cost of a terminal failure in a part of a bridge is different than in software. It is difficult to avoid juxtaposing more iterative approach conducted in the thesis and the requirement engineering approach and not to arrive results somewhat similar to comparisons of agile development methodologies and the waterfall model. One is not always superior to another, but perhaps there exist situations where one actually is.

Big Data analytics related research and practiser knowledge are presented in the thesis in chapter 2. The formed design principles incorporate some knowledge presented and discussed. Further discussion would require completion of the evaluation phase of DSRM in possible future studies.

7.2 Limitations

The largest limitation regarding the resulting design principles is the interpretation of their generalizability versus their situationality. They include and are merged with published peer-reviewed principles, include published practitioner knowledge, and include the experiences gained from technological investigations. These would contribute towards the generalizability. However, they include discovered value prioritization examined in a case project which would contribute to them being valid only situationally. To examine this issue in more detail, proper evaluation should be conducted in the future. If they are valid, are they valid only as design principles in platform build in SESP-project, valid

only on as university-based BD platform design, on research related BD platform or BD platforms overall?

Besides the scope and generalizability of the results, the validity and truthfulness of both the results and the research process should be evaluated. Results are majorly based on interpretation of values, these discovered via qualitative research which can be examined for the validity and reliability of results. Especially concern is related to the small sample of the stakeholders.

A further point of dispute for the results would be the actual implementation of the interviews. Even though prepared for with the guidance offered in literature, the inexperience in performing the interviews shows through in the resulting material. Some of the interpretations were based on answers of a single participant, thus representativeness of these interpretations could be questioned.

With the experience gained, I would re-consider the scope of the research again, as much of the demonstrative and evaluative parts of the research must be continued in possible future research. Iterative nature of the DSRM cycle is not compatible with the scope of a thesis when it is applied to a large information system such as the BD platform with many external dependencies. With the scope of a thesis, iterative loops of DSRM are more suitable to applications where prototypes and their effects can be evaluated more straightforwardly and with faster cycles of development, such as proprietary UIs.

7.3 Conclusions

Implications for practice are clear on the scale of SESP-project. Design principles identified have been used in the design and development of prototypes and continue to be used in further development. Discovered aspects exist as important limitations on design space in the project. I strongly suspect that these principles have some degree of further generalizability in the field of big data analytics, but to proving that requires further demonstration and evaluation. The approach they describe is definitely suitable for application in the area of technology that is fast evolving, dealing with the uncertainty of the future by adaptation. They point to the direction that large software projects, in the BD area, should consider carefully the time scale of the project, start and continue with evolving prototypes, integrate the users early and examine in detail the purpose of the system, and

approach practical implementation knowledge and resource requirements in the field with humility.

The implications for research are not that straightforward as for practice. Thesis proves that values can be integrated into the DSRM process systematically and as an activity conducted in the similar project space to requirement defining, leading to open and conscious discussion of values. There has been increased activity and call for value related considerations in several areas of technology and life in general. Methods of design science are applicable to several possible areas of implementation, including organizational and administrative pursuits. If it is designed, the tools of DSR are suitable. Thus, with increased interest of businesses, consumers, and citizens to reflect on values their choices exhibit, there exist a wide range of possibilities to examine the combination of different methods giving values more central role and the methods of DSR in several areas of life.

Several further possibilities for research have been previously discussed, but the most pressing would be validation on the generalizability of these principles and evaluation if a contribution to DSR knowledge has occurred or not. In practice, this could provide a challenge. The more reasonable and suitable approach would be an examination of the validity of the principles situationally, by evaluation of the continued development of the platform in the project by investigating the results of actual research usage of the platform.

REFERENCES

- Abbasi, Ahmed. Suprateek Sarker & Roger H. Chiang (2016). Big Data Research in Information Systems: Toward an Inclusive Research Agenda. *Journal of the Association for Information Systems* 17:2, 1–32.
- Acharjya, D. P. & Ahmed P. Kauser (2016). A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools. *International Journal of Advanced Computer Science and Applications* 7:2, 511–518.
- Alexander, C. (1964). *Notes on the Synthesis of Form*. 217 pp. Harvard University Press. ISBN 0-674-62751-2.
- Alshammari, Abdulwahhab & Hyunggu Jung (2017). Designing Community of Practise Systems: a Value Sensitive Approach. *Informatics, Health & Technology (ICIHT), International Conference on*. IEEE.
- Antila, E. R. Virrankoski. K. Kauhaniemi. T. Vartiainen. J. Larimo. A. Rajala. T. Galkina. S. Kock. P. Björk & A. Norrgrann (2016). *Smart Energy Systems Research Platform (SESP) Research Plan*.
- Avital, Michel. Kalle J. Lyytinen. Richard Boland Jr. Brian S. Butler. Deborah Dougherty. Matt Fineout. Wendy Jansen. Natalia Levina. Will Rifkin & John Venable (2006). Design with a Positive Lens: An Affirmative Approach to Designing Information and Organizations. *Communications of the Association for Information Systems* 18:1, Article 25.
- Apache (2018a). Apache Kafka official homepage [online]. [3.9.2018] Available: <http://kafka.apache.org/>

Apache (2018b). Apache NiFi official homepage [online]. [3.9.2018] Available:
<http://nifi.apache.org>

Apache (2018c). Apache Druid official homepage [online]. [3.9.2018] Available:
<http://druid.io/>

Apache (2018d). Apache Zeppelin official homepage [online]. [3.9.2018] Available:
<https://zeppelin.apache.org>

Apache (2018c). Apache DataFu official homepage [online]. [3.9.2018] Available:
<https://datafu.apache.org>

Begoli, Edmon & James L. Horey (2012). Design Principles for Effective Knowledge Discovery from Big Data. In: *WICSA/ECISA*, 215–218.

Briggs, Robert O. Tilo Böhmann. Gerhard Schwabe & Tuure Tuunanen (2019). Advancing Design Science Research with Solution-based Probing. *Proceedings of the 52nd Hawaii International Conference on System Sciences 2019*, 5725-5734,

Borning, Alan & Michael Muller (2012). Next Steps for Value Sensitive Design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 1125–1134.

Browne, Glenn J. (2006) Research Issues in Information Requirements Determination for Systems Development and Human-Computer Interaction. In: *Human-Computer Interaction and Management Information Systems: Applications. Advances in Management Information Systems*, Volume 6. Ed. Dennis Galletta & Ping Zhan. 480 pp. Routledge: New York. ISBN 978-0-7656-1487-2. 313–336.

Cross, N. (1993). A history of design methodology. In: *Design methodology and relationships with science*, 15–27.

- Dadgar, Majid & K.D. Joshi (2015). ICT-Enabled Self-Management of Chronic Diseases: Literature Review & Analysis Using Value-Sensitive Design. *System Sciences (HICSS)*, 2015 48th *Hawaii International Conference on*. IEEE. 3217–3226.
- Damiani, Ernesto. Claudio Ardagna. Paolo Ceravolo & Nell Scarabottolo (2017). Toward Model-Based Big Data-as-a-Service: The TOREADOR Approach. In: *Advances in Databases and Information Systems*. Springer: Cham, 2017. 3–9.
- Davenport, T.H & D.J. Patil (2012). Data Scientist: The sexiest job of the 21st century. *Harvard Business Review* 90:10, 70–76.
- Deetz, Stanley (1996). Crossroads-Describing Differences in Approaches to Organization Science: Rethinking Burrell and Morgan and Their Legacy. *Organization Science* 7:2, 191–207.
- Emani, Cheikh Kacfeh. Nadine Cullot & Christophe Nicole (2015). Understandable Big Data: A survey. *Computer Science Review* 17, 70–81.
- Fayyad, U. G. Piatetsky-Shapiro & P. Smyth (1996). Knowledge discovery and data mining: towards a unifying framework. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR*, 82–88.
- Friedman, Batya. Pether H. Kahn Jr & Alan Borning (2002). Value Sensitive Design: Theory and Methods. *University of Washington Technical Report* December 2002, 1–8.
- Friedman, Batya. Peter H. Kahn Jr & Alan Borning (2008). Value Sensitive Design and Information Systems. In: *The Handbook of Information and Computer Ethics*, 70–101. Ed. Kenneth E. Himma & Herman T. Tavani. John Wiley & Sons Incorporated. ISBN 978-047-028-1802.

- Friedman, Batya. Peter H. Kahn Jr. Alan Borning & Alina Huldtgren (2013). Value Sensitive Design and Information Systems. In: *Early Engagement and new technologies: Opening up the laboratory*, 55-96. Philosophy of Engineering and Technology, vol 16. Ed. N. Doorn, D. Schuurbiens, I. van de Poel & M. E. Gorman. Dodrecht: Springer. ISBN 978-94-007-7844-3.
- Ghemawat, Sanjay. Howard Gobioff. Shun-Tak Leung (2003). The Google file system. *SIGOPS Operating System Review*, 37:5, 29–43.
- Gregor, Shirley (2006). The Nature of Theory in Information Systems. *MIS Quarterly* 30:3, 611–642.
- Gregor, Shirley & David Jones (2007). The Anatomy of a Design Theory. *Journal of the Association of Information Systems* 8:2, 313–335.
- Gregor, Shirley & Alan R. Hevner (2013). Positioning and Presenting Design Science Research for Maximum Impact. *MIS Quarterly* 37:2, 337–355.
- Hashem, I.A.T. I. Yaqoop. N.B. Anuar. S. Mokhtar. A. Gani & S.U. Khan (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems* 47, 98–115.
- Hevner, Alan R (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*. 19:2, Article 4.
- Huppertz, DJ (2015). Revisiting Herbert Simon's "Science of Design". *Design Issues* 32:2, 29–40.
- Hoven van den, Jeroen (2013). Value Sensitive Design and Responsible Innovation. In: *Managing the Responsible Emergence of Science and Innovation in Society*, 76–

83. Ed. Richard Owen. John Bessant & Maggy Heintz. John Wiley & Sons Incorporated. ISBN 978-111-855-141-7.

Hsieh, Hsiu-Fang & Sarah E. Shannon (2005). *Three Approaches to Qualitative Content Analysis*. *Qualitative Health Research*, 15:9.

Iivari, Juhani (2007). A Paradigmatic Analysis of Information Systems As a Design Science. *Scandinavian Journal of Information Systems* 19:2, Article 5.

Jacobs, Naomi & Alina Huldtgren (2018). Why value sensitive design needs ethical commitments. *Ethics and Information Technology*, In Press. 1–4.

Johri, Aditya & Sumitra Nair (2011). The role of design values in information system development for human benefit. *Information Technology & People*, 24:3. 281–302.

Järvinen, Pertti (2017). Two Different Goals in Design Science Research, One from Science Another from Practise. University of Tampere. *Reports in Information Sciences* 2017:52, ISBN 978-952-03-0401-0 (pdf).

Koskimies, Kai & Tommi Mikkonen (2005). *Ohjelmistoarkkitehtuurit*. Helsinki: Talentum. ISBN 952-14-0862-6.

Kurgan, Lukasz A. & Petr Musilek (2006). A survey of Knowledge Discovery and Data Mining process Models. *The Knowledge Engineering Review* 21:1, 1–24.

Laplante, Philip A. (2014). *Requirements Engineering for Software and Systems*. 2nd Ed. Boca Raton: Taylor & Francis Group. 302 pp. ISBN 978-1-4665-6081-9.

MacGregor, John. (2013). *Predictive Analysis with SAP: The Comprehensive Guide*. SAP Press, 2013.

- Manders-Huits, Noëmi (2011). What Values in Design? The Challenge of Incorporating Moral Values into Design. *Science and Engineering Ethics* 17:2, 271–287.
- Mazunder, S. (2016). Big Data Tools and Platforms. In: *Big Data Concepts, Theories and Applications*. Ed. Shui Yu & Song Guo. Switzerland: Springer International Publishing. 29–129. ISBN 978-3-319-27763-9.
- Mendelevitch, Ofer. Casey Stella & Douglas Eadline (2017). *Practical Data Science with Hadoop and Spark*. Addison-Wesley. 230 pp. ISBN 978-0-13-402414-1.
- Miller, Jessica. Batya Friedman. Gavin Jancke & Brian Gill (2007). Value tensions in design: the value sensitive design, development and appropriation of a corporation's groupware system. *Proceedings of the 2007 international CAM conference on Supporting group work*. ACM. 281–290.
- Mok, Luisa & Sampsa Hyysalo (2018). Designing for energy transition through Value Sensitive Design. *Design Studies* 54, 162–183.
- Moreno-Garcia, I. M. A. Moreno-Munoz. V. Pallas-Lopez. M.J. Gonzalez-Redondo. E. J. Palacios-Garcia & C.D. Moreno-Moreno (2017). Development and application of a smart grid test bench. *Journal of Cleaner Production* 162, 45–60.
- Lee, Allen S. Manoj Thomas & Richard L Baskerville (2015). Going Back to Basics in Design Science: from the Information Technology Artifact to the Information Systems Artifact. *Information Systems Journal*, 25:1, 5–21.
- Oussous, Ahmed. Fatima-Zahra Benjelloun. Ayoub Ait Lachen & Samir Belfikh (2017). Big Data Technologies: A Survey. *Journal of King Saud University - Computer and Information Sciences* (2017). In Press.

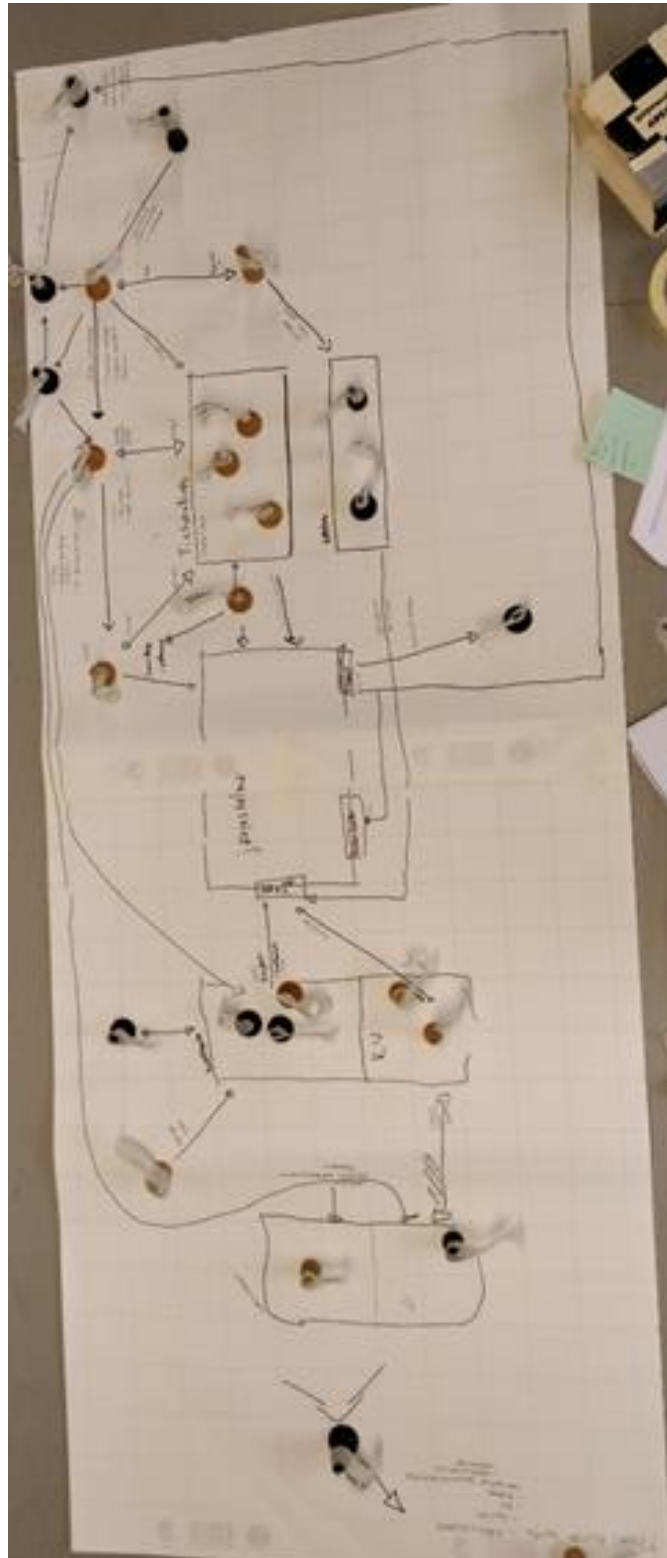
- Oxford (2019). Oxford Advanced Learner's Dictionary [online]. [17.1.2019]. Available: <https://www.oxfordlearnersdictionaries.com/definition/english/>
- Peppers Ken. Tuure Tuunanen. Marcus A. Rothenberger & Samir Chatterjee (2008). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems* 24:3, 35–77.
- Pushshift (2019). [online] An online dataset of Reddit messages. Available: <https://files.pushshift.io/>
- Pommeranz, Alina. Christian Detweiler. Pascal Wiggers & Catholijn Jonker (2012). Elicitation of situated values: need for tools to help stakeholders and designers to reflect and communicate. *Ethics and Information Technology* 14:4, 285–303.
- Roman, Javi (2018). Hadoop Ecosystem Table [online]. [10.1.2019] Available: <https://hadoopecosystemtable.github.io/>
- Rowley, Jennifer (2012). Conducting research interviews. *Management Research Review* 35:3, 260–271.
- Serrador, Pedro & Jeffrey K. Pinto (2015). Does Agile work? – A quantitative analysis of agile project success. *International Journal of Project Management*, 33:5, 1040–1051.
- Simon, Herbert A. (1996). *The Sciences of the Artificial*. 3rd Ed. 4th Printing. United States of America: Massachusetts Institute of Technology. 231 pp. ISBN 0-262-19374-4.
- Sonnenberg, Christian & vom Brocke, Jan (2008). Evaluation Patterns for Design Science Research Artefacts. In: *Proceedings of the European Design Science Symposium (EDSS) 2011*. Ed. M. Helfert & B. Donnellan. Springer: Dublin.

- Tiainen, Tarja (2014). Haastattelu tietojenkäsittelytieteen tutkimuksessa. *Informaatiotieteiden yksikön raportteja 25/2014*. Tampereen yliopisto. ISBN 978-951-44-9374-4.
- Yoo, Daisy (2018). Stakeholder Tokens: a constructive method for value sensitive design stakeholder analysis. *Ethics and Information Technology*, In Print, 1–5.
- Walls, Joseph. George Widmeyer & Omar Sawy (1992). Building an Information System Design Theory for Vigilant EIS, *Information Systems Research* 3:1, 36–59.
- Wang, Hai & Zeshui Xu (2016). Towards Felicitous Decision Making: an Overview on Challenges and Trends of Big Data. *Information Sciences* 367, 747–765.
- White, Tom (2015). *Hadoop: The Definitive Guide. Storage and Analysis at Internet Scale*. Fourth Ed. Sebastopol: O'Reilly Media. 727 pp. ISBN: 978-1-491-90163-2.
- Wiig, K. M. (1993). *Knowledge Management Foundations: thinking about—how people and organizations create, represent, and use knowledge*. Arlington, Texas: Schema. 474 pp. ISBN 0-9638925-0-9.
- VMware (2018). *VMware ESXi Installation and Setup*. Online pdf-formatted instruction manual. Available: <https://docs.vmware.com/en/VMware-vSphere/6.7/vsphere-esxi-67-installation-setup-guide.pdf>
- Wynsberghe, Aimee (2013). Designing Robots for Care: Care Centered Value–Sensitive Design. *Science and Engineering Ethics* 19:2, 407–433.
- Xu, Heng. Robert E. Crossler & France Bélanger (2012). A Value Sensitive Design Investigation of Privacy Enhancing Tools in Web Browsers. *Decision Support Systems* 54, 424–433.

Zhang, Yaoxue. Ju Ren. Jiagang Liu. Chugui Xu. Hui Guo & Yaping Liu (2017). A Survey on Emerging Computing Paradings for Big Data. *Chinese Journal of Electronics* 26:1, 1–12.

APPENDICES

APPENDIX 1. Results of the stakeholder mapping session.



APPENDIX 2. Survey Questions in Finnish.

Haastattelulomake 1.2, SESP WP6

Päivämäärä: _____

Nimi : _____

Syntymävuosi : _____

Ammatti : _____

SESP-hankkeessa rakennetaan analyysi- ja tallennusjärjestelmää monenlaisen, monimuotoisen ja kooltaan vaihtelevan tiedon analysointia ja käsittelyä varten. Järjestelmä itsessään mahdollistaisi monimuotoista käyttöä ja hyödyntämistä yliopistoyhteisössä sekä erilaisten yhteyksien rakentamisen. Tässä kartoitetaan **mitä järjestelmää joko suorasti tai epäsuorasti käyttävät, tai järjestelmään muuten vaikuttavat henkilöt, pitävät tärkeinä asioina data-alustaan liittyen idea- että arvotasolla.**

Esitettyyn kuvaan liittyen ja järjestelmän havainnollistamiseksi. Tarkoitus ei ole vastata näihin kysymyksiin, vain hieman havainnollistaa järjestelmää ja sen mahdollisuuksia.

1. Onko olemassa yliopiston lisäksi muita relevantteja järjestelmään mahdollisesti liittyviä toimijoita – yhteisöjä, instituutioita, yrityksiä tai tutkimuslaitoksia?
2. Jotka vaikuttavat yliopistoon ja/tai ovat vuorovaikutuksessa yliopiston kanssa jollain tavalla?
3. Tai osallistuvat tai vaikuttavat järjestelmän käyttöön?
4. Käyttävät tai hyödyntävät järjestelmää suoraan?
5. Käyttävät, hyödyntävät, ostavat, jatkojalostavat tai muuten ovat tekemisissä erilaisten järjestelmän tuottamien asioiden kanssa?
6. Mikä itse käyttämisessä tai hyödyntämisessä on mielestäsi tärkeää?

Teema 1. Järjestelmä kokonaisuutena ja sen elinkaari.

Kysymys 1: Millaisen järjestelmän tulisi olla kokonaisuudessaan valmiina mielestäsi?

Kysymys 2: Millaisena näkisit järjestelmän elinkaaren? Mistä se lähtee, mitä elinkaaren aikana tapahtuu, mihin se päättyy?

Kysymys 3: Edelliseen liittyen, minkälaisia haittoja järjestelmään liittyen pystyt kuvittelemaan? Nämä voivat liittyä esimerkiksi suoraan käyttöön, käyttämisen tulokseen, erilaisten käyttäjien tai yhteisöjen keskinäisiin suhteisiin tai ristiriitoihin?

Teema 2. Käyttö ja käyttäjät.

Kysymys 4: Mihin järjestelmää mielestäsi tulisi käyttää?

Kysymys 5: Keiden kaikkien sitä tulisi käyttää?

Kysymys 6: Mikä näistä mainitsemistasi käytöistä on tärkeintä? Miksi?

Teema 3. Oma käyttö.

Kysymys 7: Haluaisitko itse hyödyntää järjestelmää jollain tavalla? Jos, niin miten? Jos ei, miksi ei?

Kysymys 8a. Mikäli haluaisit hyödyntää järjestelmää jollain tavalla, mikä mainitsemassasi käytössä on oleellista ja mielestäsi tärkeätä? Kenties erikoista tai poikkeavaa?

8b. Tietoaineiston kannalta? Tiedon muodossa tai määrässä? Nopeudessa? Tiedon oikeudelliset rajoitteet? Kuinka vaihtelevia aineistoja? Tietoaineiston omistaminen?

8c. Analyysivaiheessa. Koneoppimisen hyödyntämistä? Laadullista vai määrällistä analyysiä?

8d. Tulosten kannalta? Tulosten omistaminen ja jakaminen? Kaupallinen hyödyntäminen? Yhteistyössä laadittuja?

8e. Muu mahdollinen suora tai epäsuora järjestelmän hyödyntäminen. Mikä on mahdollisessa muussa hyödyntämisessä mielestäsi keskeistä?

Teema 4. Miten näkisit seuraavien asioiden tärkeysjärjestyksen järjestelmän suhteen? Numeroi **viisi** tärkeintä: 1 merkkää tärkeintä, 2 seuraavaksi tärkeintä jne. Kun numeroit, voitko kertoa ääneen mitä tarkoitat kohdalla. Tyhjiin kohtiin voi kirjoittaa oman.

- Avoimuus
- Edullisuus
- Jaettavuus
- Kehitettävyyys
- Käytettävyyys
- Liitettävyyys
- Monikäyttöisyys
- Opetus
- Tietoaineistopankki
- Tulosten kiinnostavuus, tieteellinen
- Tulosten kiinnostavuus, kaupallinen
- Turvallisuus
- Tutkimus
- Yhteistyö, kansainvälinen
- Yhteistyö, paikallinen
- Yhteistyö, kaupallinen
- Yhteistyö, tieteellinen
- Yksityisyys
- Ymmärrettävyyys
- _____
- _____

APPENDIX 3. Survey Questions in English.

Interview form 1.2, SESP WP6

Date: _____

Name : _____
 Year of Birth : _____
 Occupation : _____

One of the main objectives of the SESP–project is to create a data analysis and storage platform which is suitable for processing, analyzing and storing wide variety of data. Platform should be capable of accommodating a multitude of use cases in university context and allow implementation of different connections. With this survey is investigated **what direct and indirect identified stakeholders of the system consider important in the data platform context, both in idea and value level.**

Purpose of the next questions is twofold, to offer a bit of warmup and to offer context for participants who have not necessarily thought about the system before.

1. Do there exist other relevant entities related to the system besides University? Communities, institutions, companies or research institutions?
2. That influence the university and/or interact with it?
3. Participate or affect the actual usage of the system?
4. Make use of the system directly?
5. Use, buy, refine or otherwise are interacting with different outputs of the system?
6. What is important in using or in making use of the system, in your mind?

Theme 1. System in entirety and the lifecycle of the system.

Question 1: How do you see the entirety of the system?

Question 2: How do you see the lifecycle of the system? Where does it start, what happens during the lifecycle and how does it end?

Question 3: What kind of harms or disadvantages related to the system can you imagine? These can be related to direct usage of the system, results of usage, relations or conflicts between different users, user groups or related entities?

Theme 2. Using of the system and the users.

Question 4: What the system should be used for?

Question 5: Who should use the system?

Question 6: What use would you consider the most important?

Theme 3. Personal use of the system.

Question 7: Would you personally make use of the system in some way? Direct or indirect? If so, how? If not, why not?

Question 8a. If you would like to make use of the system in some way, what would be most substantive in your use and what would consider most important in it? Perhaps something special or different?

8b. Regarding your data? In the format or amount of the data? Velocity of the data? Legal or contractual restrictions regarding data? Ownership of the data?

8c. In the analysis and processing phase. Application of for example machine learning? Qualitative or quantitative techniques?

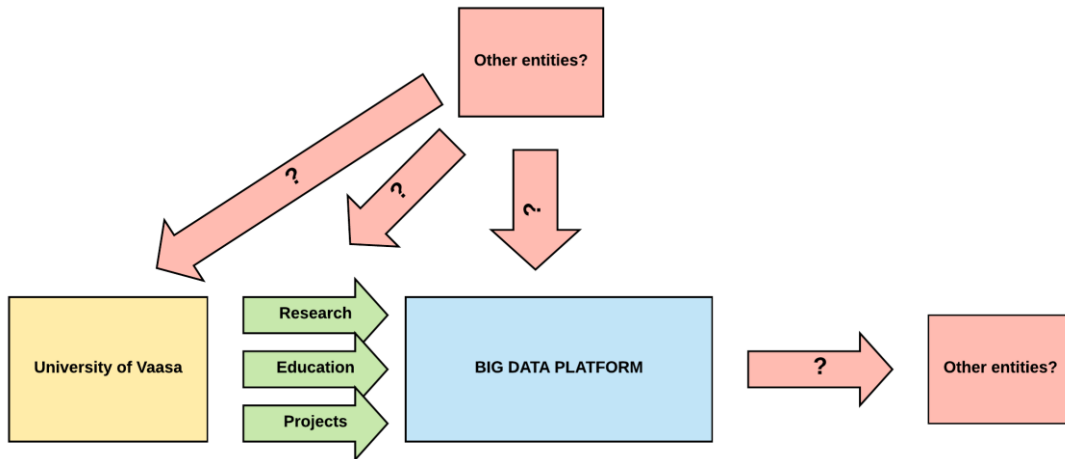
8d. Regarding results? Ownership of results and sharing of them? Commercial usage? Results reached by different means of co-operation?

8e. Other direct or indirect utilization of the system. What would you consider central or essential in other utilization of the system?

Theme 4. How would you prioritize the following things in the context of the system? Would please assign numbers from 1 to 5 to each, 1 marking the most important thing, 2 the second-most important and so on. As you assign numbers, would you please think out aloud what do you mean with the concept. At the end of the list are two empty fields to fill out in the case you think a central concept is missing.

- Affordability
- Connectivity
- Co-operation, Scientific
- Co-operation, Commercial
- Co-operation, International
- Co-operation, Local
- Data storage
- Developability
- Distributability, shareability
- Interestingness of results, scientific
- Interestingness of results, commercial
- Openness
- Privacy
- Research
- Security
- Teaching
- Understandability
- Usability
- Versatility

APPENDIX 4. Interview Warm-up Diagram.



APPENDIX 5. Full Result Table of Theme 4.

Concept	Points
Research	38
Usability	36
Security	22
Versatility	19
Developability	16
Teaching	15
Data Storage	12
Openness	11
Connectivity	11
Co-operation, commercial	11
Co-operation, scientific	10
Privacy	10
Interestingness of results, scientific	9
Interestingness of results, commercial	7
Understandability	5
Distributability, shareability	3
Co-operation, International	3
Affordability	1
Ensuring of the continuity of system	1
Co-operation, local	0

APPENDIX 6. The component stack and initial versions in prototype 2.

Component	Version
HDFS	3.1.1
YARN	3.1.1
MapReduce2	3.1.1
Tez	0.9.1
Hive	3.10
HBase	2.0.0
Pig	0.16.0
Sqoop	1.4.7
Oozie	4.3.1
ZooKeeper	3.4.6
Accumulo	1.7.0
Infra Solr	0.1.0 (not activated)
Ambari Metrics	0.1.0 (not activated)
Atlas	1.0.0
Kafka	1.1.1
Knox	1.0.0
SmartSense	1.5.0.2.7.1.0-169 (not activated)
Spark2	2.3.1
Zeppelin Notebook	0.8.0
Druir	0.12.1

APPENDIX 7. Component distributions in the cluster.


Master 1 node:

Status	Name	Type	Action
	Accumulo Master / Accumulo	Master	...
	Accumulo Tracer / Accumulo	Master	...
	Timeline Service V1.5 / YARN	Master	...
	Druid Coordinator / Druid	Master	...
	Druid Overlord / Druid	Master	...
	History Server / MapReduce2	Master	...
	HST Server / SmartSense	Master	...
	Infra Solr Instance / Infra Solr	Master	...
	Kafka Broker / Kafka	Master	...
	Knox Gateway / Knox	Master	...
	Metrics Collector / Ambari Metrics	Master	...
	NameNode / HDFS	Master	...
	ResourceManager / YARN	Master	...
	YARN Registry DNS / YARN	Master	...
	ZooKeeper Server / ZooKeeper	Master	...
	HST Agent / SmartSense	Slave	...
	Metrics Monitor / Ambari Metrics	Slave	...
	HDFS Client / HDFS	Client	...
	Infra Solr Client / Infra Solr	Client	...
	Tez Client / Tez	Client	...

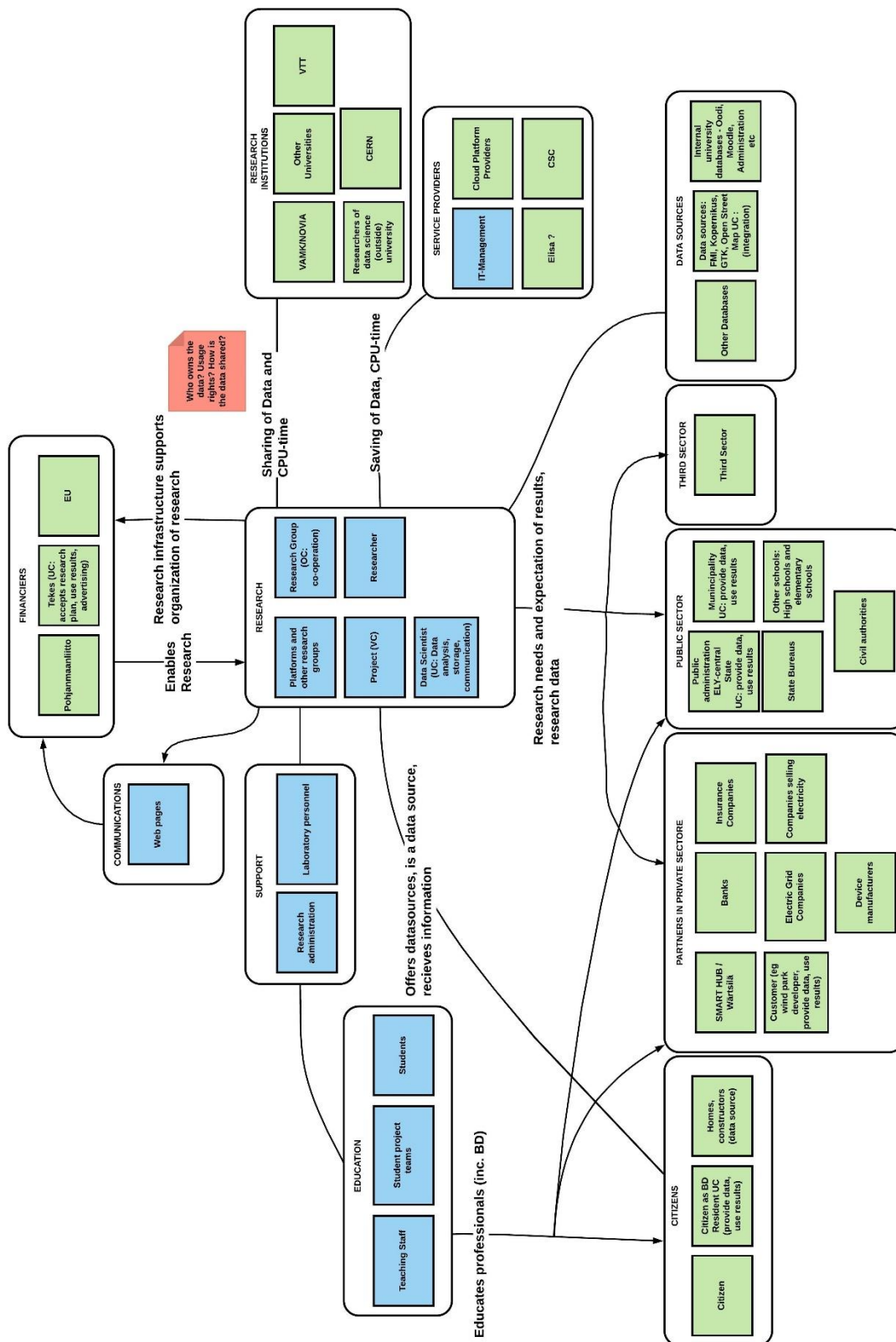
Master 2 node:

Status	Name	Type	Action
	Accumulo GC / Accumulo	Master	...
	Accumulo Monitor / Accumulo	Master	...
	Activity Analyzer / SmartSense	Master	...
	Activity Explorer / SmartSense	Master	...
	Atlas Metadata Server / Atlas	Master	...
	Druid Broker / Druid	Master	...
	Druid Router / Druid	Master	...
	Active HBase Master / HBase	Master	...
	Hive Metastore / Hive	Master	...
	HiveServer2 / Hive	Master	...
	Grafana / Ambari Metrics	Master	...
	MySQL Server / Hive	Master	...
	Oozie Server / Oozie	Master	...
	SNameNode / HDFS	Master	...
	Spark2 History Server / Spark2	Master	...
	Timeline Service V2.0 Reader / YARN	Master	...
	ZooKeeper Server / ZooKeeper	Master	...
	HST Agent / SmartSense	Slave	...
	Metrics Monitor / Ambari Metrics	Slave	...
	HBase Client / HBase	Client	...
	HDFS Client / HDFS	Client	...
	Hive Client / Hive	Client	...
	Infra Solr Client / Infra Solr	Client	...
	MapReduce2 Client / MapReduce2	Client	...
	Tez Client / Tez	Client	...
	YARN Client / YARN	Client	...
	ZooKeeper Client / ZooKeeper	Client	...

Example of slave node, notepad.uwasa.fi:

Status	Name	Type	Action
✓	Zeppelin Notebook / Zeppelin No...	Master	...
✓	ZooKeeper Server / ZooKeeper	Master	...
✓	Accumulo TServer / Accumulo	Slave	...
✓	DataNode / HDFS	Slave	...
✓	Druid Historical / Druid	Slave	...
✓	Druid MiddleManager / Druid	Slave	...
✓	RegionServer / HBase	Slave	...
✓ 	HST Agent / SmartSense	Slave	...
✓	Metrics Monitor / Ambari Metrics	Slave	...
✓	NodeManager / YARN	Slave	...
✓	Accumulo Client / Accumulo	Client	...
✓	Atlas Metadata Client / Atlas	Client	...
✓	HBase Client / HBase	Client	...
✓	HDFS Client / HDFS	Client	...
✓	Hive Client / Hive	Client	...
✓	Infra Solr Client / Infra Solr	Client	...
✓	MapReduce2 Client / MapReduce2	Client	...
✓	Oozie Client / Oozie	Client	...
✓	Pig Client / Pig	Client	...
✓	Spark2 Client / Spark2	Client	...
✓	Sqoop Client / Sqoop	Client	...
✓	Tez Client / Tez	Client	...
✓	YARN Client / YARN	Client	...
✓	ZooKeeper Client / ZooKeeper	Client	...

APPENDIX 8. Workshop result of Team one.



APPENDIX 9. Workshop result by Team two

