BACHELOR THESIS

# Compliance of Data Lake Enterprise Architecture Model with the General Data Protection Regulation (GDPR)

Vuk Kadenic

2015

Bachelor of Arts
Systems Science

Luleå University of Technology
Department of Computer science, Electrical and Space engineering

# Preface

This report details my Bachelor Thesis of the System Science program at Luleå University of Technology (LTU). The Thesis was done on behalf of Capgemini, renowned consulting company with offices in Stockholm, Sweden. The idea of writing a thesis on this particular subject was presented to me by Christofer Holmgren Bagge, Managing Consultant and Senior Business Analyst with focus on Big Data and Big Data Analytics at Capgemini Sweden, who is currently involved in several major compliance- and governance-related projects.

I would like to thank Capgemini and Christofer Holmgren Bagge in particular for giving me the opportunity to conduct this work and for their help in accomplishing it. Christofer was always there for me when I needed help with gathering information or to discuss new ideas. His feedback was much appreciated and has helped me stay on the right path throughout the thesis.

I would also like to thank the staff and my colleagues at LTU for their feedback and help with both the professional and the academic aspect of this work. I would especially like to thank my friend and colleague Jari Kaljunen for his valuable feedback and inspired discussions.

Finally, I would like to thank my family for being there for me during my studies. To my beautiful son Nikola and loving daughter Silvia, I lost many precious moments with you that I now promise to make up. To my dear wife Nikolina I owe a very special debt for her unwavering love, loyalty, honor and support in my moments of need. Without her inspiration, the road that led to this thesis would not have been as memorable.

# Abstract

The purpose of this thesis is to determine if Data Lake Enterprise Architecture model is compliant with the coming EU's General Data Protection Regulation (GDPR) and to suggest the design of eventual changes that would improve the model towards reaching compliance. While Capgemini-Pivotal's Business Data Lake was used as a reference Data Lake model, results of this thesis are applicable to Data Lake architectures in general. It can be argued that even other solutions for data management in Big Data environment will benefit from the proposed design for management of personal data.

During the course of this work, baseline architecture of Business Data Lake was captured on all three levels of Enterprise Architecture, target architecture was developed and gaps were identified. The gap analysis effectively revealed the issues that baseline architecture has in achieving the intended compliance and has led to the selection of a key gap that was bridged through the proposed design of a dedicated repository for personal metadata. The design itself is based on a graph database that in a revolutionary way improves the performance of personal data management. Finally, the solution design is demonstrated through a working prototype of graph-based repository for personal metadata.

The thesis and its results were evaluated by Capgemini Sweden as the project's sponsor and were largely found to fulfill and in some areas even exceed the goals set before this project. Other, important aspects of Business Data Lake's compliance with GDPR, which are not addressed by this work, are also explained.

It was concluded that thesis accomplished its purpose in determining that the Data Lake architecture model, in its current state, fails to comply with the GDPR. Moreover, gaps that need to be bridged in order for compliance to be reached were identified and a key solution was proposed and prototyped. It was also concluded that the thesis represents a good base for continuance of the design work, and that further design projects are necessary if the compliance is to be achieved.

# Sammanfattning

Syftet med denna avhandling är att avgöra om Data Lake Enterprise Arkitektur modell är förenlig med den kommande EU:s allmänna dataskydds förordning (General Data Protection Regulation - GDPR), samt att föreslå utformningen av eventuella förändringar som skulle förbättra modellens kompatibilitet med förordningen. Även om Capgemini-Pivotal:s Business Data Lake användes som referent Data Lake modell, resultaten av denna avhandling är tillämpliga på Data Lake arkitekturer i allmänhet. Det kan även argumenteras att andra lösningar för data hantering i en Big Data miljö kommer att ha nytta av den föreslagna designen för hantering av personlig data.

Under arbetets gång har nulägets arkitektur av Business Data Lake kartlagts på samtliga tre nivåer av Enterprise Arkitekturen, för att vidare fortsätta med utvecklingen av målarkitekturen och genomförandet av gapanalys som avslöjade klyftor (gaps) mellan två arkitekturslägen. Gapanalysen effektivt påvisade de problemen som nulägets arkitektur har med att uppnå förenlighet med GDPR och har lett till urvalet av en nyckel gap som överbryggdes genom den föreslagna utformningen av en särskild lagringsplats för personlig metadata. Själva designen är baserad på en grafdatabas som på ett revolutionärt sätt förbättrar prestanda av hanteringen av personlig data. Slutligen, den föreslagna lösningen demonstrerades genom en fungerande prototyp av en grafdatabas-baserat lagringsplats för personlig metadata.

Avhandlingen och dess resultat utvärderades av Capgemini Sverige som projektets beställare och har bedömds att till en stor del uppfyller och i vissa områden till och med överstiger de uppsatta målen för projektet. Andra, viktiga aspekter av Data Lake:s kompatibilitet med GDPR, som inte tas upp i avhandlingen, har också förklarats.

Det konstaterades i slutsatsen att avhandlingen har uppnått sitt syfte genom att fastställa att Data Lake arkitektursmodell, i dess nuvarande skick, avviker från GDPR. Dessutom, klyftor (gaps) som behöver överbryggas för att avvikelser skall upphöras har identifierats och lösningen till en nyckel gap föreslogs och gjordes en prototyp av.  I slutsatsen vidare konstaterades att avhandlingen utgör en bra bas för fortsättningen av designarbetet, samt att ytterligare designprojekt är nödvändiga för att förenlighet av Data Lake Enterprise Arkitektur modell med GDPR skall kunna uppnås.

# Table of Contents

# 1 Introduction

Big Data has become one of the most important technology trends of Business Intelligence, enhancing the ability to create competitive advantage by giving deeper analytical insights and accelerating analytical process as a whole (Zikopoulos et al., 2012).

The road to Big Data was paved by drastic growth of collected datasets coming from new sources of data such as sensors, social media and website interactions. Complementing the growth problem, collected data comes in different form and it varies both in type and timelines. Traditional data management solutions, such as data warehouses and data marts, were designed to deal with large amounts of structured data fed in batch intervals, usually weekly or daily. For large amounts of semi-structured and unstructured data, as well as tracking changes in real-time, the designed proved inadequate. For a time, ODBMS (Object Data Management Systems) seemed to be the solution, but new requirements, such as even larger amounts and varieties of data, data virtualization and cloud computing created the need for a new approach. This approach was called Big Data and embodies the combination of different technologies, such as virtualization, parallel processing, distributed file systems, advanced analytics, MapReduce and Hadoop (Zikopoulos et al., 2012). Ohlhorst (2013), describes Big Data as a concept that has evolved from a situation in which data sets have grown beyond the ability of conventional information technologies to effectively manage either the size or the scale and growth of the data set. The concept evolved to include not only the size of data sets, but also the processes involved in managing these data sets. Because of that, the concept of Big Data has even become interlaced with concepts of business intelligence, analytics, and data mining (Ohlhorst, 2013). Zikopoulos et al., (2012), summarizes the advantages of Big Data approach in that it enables gaining valuable insights from many different types of data at the right speed, regardless of data amount.

There are many different implementations of the Big Data approach, but this thesis concentrates on a Data Lake model. According to Oliver (2014), Data Lake model can, in a very simplified manner, be described as an unstructured data warehouse, where all the data is cast in one large pool without cleaning and structuring it. The main advantage of this approach is that the said data is available on an enterprise level in its native, raw form, thus eliminating the need for integration between different departments. Previous approaches, that structured the data before sharing it, constrained the use of data because when data is cleaned and structured for use by one department, it can lose the properties of importance to another. Using Data Lake approach preserves the value of data regardless of intended purpose. Instead of structuring data on Write, when it is extracted into Data Lake, data is structured on Read, responding to user's query.

According to Tene and Polonetsky (2012), data that is being amassed and analyzed creates enormous value for global economy, as it drives innovation, productivity, efficiency, and growth. At the same time they predict that privacy concerns could lead to "regulatory backlash, dampening the data economy and stifling innovation" (Tene and Polonetsky, 2012). These privacy concerns were well founded. According to Ohm (2010), the advances of computer science, namely Big Data technologies, have rendered anonymization of Personal Identifiable Information useless. Ohm's claim was confirmed in practice by the case of an anonymized Netflix dataset, containing user's comments, being comprised through external correlation with Internet Movie Database (Narayanan & Shmatikov, 2008).

The prediction about regulatory backlash became reality in the form of General Data Protection Regulation (referred to as GDPR in further text), which was proposed in 2012 and was adopted as draft by European Parliament in 2014. Although the final technical details of the law are still being developed at the time of this writing, the law's principles are agreed on by both the EU Parliament and the EU Council and the law is expected to be finalized 2015/2016 and to come into force 2017/2018 (European Commission, 2014).

Capgemini has, through cooperation with Pivotal, become one of the leading vendors offering Big Data solution based on a Data Lake model. I was proposed by Christofer Holmgren Bagge, Managing Consultant and Senior Business Analyst with focus on Big Data and Big Data Analytics at Capgemini Sweden, to explore the compliance of Data Lake model with GDPR as my bachelor thesis.

## 1.1 Problem

According to Davis, M.'s(2014) summary of GDPR proposal that was adopted by EU Parliament 2014/03/12 all organizations processing the data concerning EU citizens, regardless of where the data is stored, would have to implement the following points:

1. Collect explicit consent to collect data from data subjects and provide the option of withdrawing that consent.
2. Facilitate "Right to erasure" – provide a service of deleting all data concerning the data subject upon request, unless there is a legitimate reason to deny said request.
3. Establish a clear privacy policy that will be provided to all data subjects.
4. Upon request, provide data subjects with a copy of their personal data in a format that can be transmitted electronically to another system.
5. Undertake annual risk management/analysis.
6. Decide on which is to be Single Data Protection Authority (DPA) for the organization.
7. Appoint a Data Controller inside the organization that would be responsible for compliance of all data processing operations with GDPR.
8. Appoint a lead authority Data Controller that would be responsible for all data processing operations across Europe.
9. Organizations processing data of more than 5000 data subjects would have to appoint a Data Protection Officer, responsible for organization's compliance with GDPR.

The proposition also includes joint liability of Data Controller and Data Processor (even the cloud provider) for any breach (Davis, M., 2014).

How compliance requirements affect Big Data is generally harder to understand than in the case of traditional enterprise database server environment, according to Ohlhorst, (2013). Ohlhorst further elaborates that new data types and methodologies are still expected to meet the legislative requirements placed on businesses by compliance laws and that there will be no excuses accepted and no passes given if a new data methodology breaks the law.

Data Lake model is built on a principle that the structure of data is defined at the time the data is used. This principle is referred to as schema on read and it enables capturing and storing raw data at scale for a low cost (CITO Research, 2014). To implement certain points of GDPR, such as Right to erasure and providing data subjects with a copy of their personal data, it is necessary to be able to identify the personal data at the time of data’s ingestion into the lake. Whether or not such functionality exists today for all data sources and data types is unclear, but even if some form of personal data identification would be proved to exist today, GDPR broadens the definition of Personal Data to include any information relating to an individual, whether it relates to his or her private, professional or public life (Triger, 2014). Trigger further elaborates that Personal Data can be anything from a name, a photo, an email address, bank details, postings on social networking websites, medical information or a computer's IP address. This effectively makes any existing solutions for data ingestion obsolete in relation to GDPR.

Tene and Polonetsky, (2012, p.67), state that: ”Privacy and data protection laws are premised on individual control over information and on principles such as data minimization and purpose limitation.”. This statement places GDPR on a direct collision course to Big Data approach in general, but especially to Data Lake model that, according to Harper (2014), delivers access to all data throughout the enterprise and that stores semi-structured and unstructured data in which initial data attributes have yet to be determined. The seriousness of this problem is recognized by Harper in a conclusion that: “The future of Data Lakes depends on their capacity to reinforce governance and address inherent security issues.” (Harper, 2014).

Whether or not Data Lake model is compliant with GDPR, to what extent, and what changes would be necessary to achieve the compliance has never been evaluated, until now. That fact may not be so surprising, according to Ipswitch’s European online survey conducted in October 2014. The survey was conducted on 316 IT professionals across Europe and the results showed that 56% of the respondents could not accurately identify what GDPR means. 52% openly admitted that they are not ready for GDPR, 35% responded that they don’t know whether their policies and processes would suffice, while only 12% responded that they feel ready for the coming changes (Ipswitch, 2014).

The penalties proposed for failing to enforce GDPR that go up to 5% global turnover of the whole organization, up to the maximum of 100000000 (one hundred million) EUR, underline the seriousness of this problem (European Parliament, 2014).

As far as implementation goes, some of the listed points of GDPR may prove to be less problematic to incorporate into Data Lake model than the others. There are however points that may well be impossible to implement due to technical constraints. One such point may prove to be “Right to erasure”, due to the fact that Hadoop file system is immutable and that files are not being physically updated or deleted (White, 2012). Old versions of files remain in Hadoop file system and, while access to them can be regulated, it is an open question whether this solution is in accordance with GDPR. Answers to such questions may come through further refinement of GDPR. However, at this time a reservation will be noted for all compliance issues that remain potentially unsolved due to such constraints.

## 1.2 Goals

The main goal of this thesis is to investigate the compliance of Data Lake architecture model with GDPR and to design the necessary changes in order to make it compliant. Even though a baseline architecture model of Capgemini-Pivotal’s Business Data Lake will be used, the results of this thesis should be applicable to Data Lake architecture model in general due to the fact that all Data Lake models are based on the same open source technology (CITO Research, 2014). The resulting architecture model shall also be based on open source technology, encouraging further development. The goal will be accomplished in following steps:

1. Baseline architecture of Capgemini-Pivotal's Business Data Lake will be mapped.
2. Target architecture will be designed according to requirements derived from GDPR.
3. GAP analysis will be performed.
4. Optionally, a prototype of the solution will be designed.
5. The results will be discussed with and evaluated by Capgemini.

The secondary goal of the thesis is to shed light on changes that GDPR shall soon enforce upon management of personal data and Big Data in general. Oracle (2015), states that 90% of world's data was created during the last two years, while Jonatan Shaw (2014) of Harward Magazine concurs in that the total data accumulation of the past two years dwarfs the prior record of human civilization. These statements combined with the result of Ipswitch's (2014) study about GDPR awareness imply the importance of this goal as well.

The scientific method used for reaching abovementioned goals and to answer the research questions is design research, which implies that new knowledge shall be gained through designing the target architecture, subsequent GAP analysis, and solution prototype.

As a design method, TOGAF ADM 9.1 will be used, with Archimate 2.1 as a description language.

## 1.3 Research questions

The research questions this thesis aspires to answer are:

I. Is Data Lake Enterprise Architecture model compliant with GDPR and to what extent?
II. What changes would be necessary to design in order achieve the compliance?

## 1.4 Limitations

- The existing policies in Business Data Lake are presumed to be in compliance with current EU legislation – Data Protection Directive. Because GDPR does not directly imply significant changes on the policy level, the compliance of policies will not be the focus of this thesis.
- Planning and managing the change of current enterprise governance and support models is a process that could really be started only after validation of produced artefacts and is therefore out of scope of this work. Some considerations about some aspects of this process may, however, be given in order to help future development.
- The scope of this work is limited to Enterprise Architecture model, no deployed instance of Business Data Lake Enterprise Architecture is the subject of this Thesis.
- Enterprise architecture views will be designed on a relatively high abstraction level in order to keep the focus on the main changes of the architecture that are relevant to compliance with GDPR.
- Hadoop immutability constraint is an issue that is out of the scope of this work. The reason behind this decision is that this is a huge technology issue, and as such it would be presumptuous to attempt its solution as one of the requirements of this project.

- For the similar reason, regulation of the ingestion process in a way that incoming data is reliably screened for personal data is out of the scope of this work. This process should be designed in a way that it includes sources of data, and could require strict standardization of data extraction and ingestion, in terms of metadata generation.

# 2 Theoretical background

Information Technology is a fast changing field, while the subject of this thesis is extremely current. Usually this is a recipe for a sparse theoretical background. However, the importance of the subject and the fact that Big Data is perhaps the most dominant trend of Business Intelligence today has given the abundance of sources available. The theoretical background for this thesis is formed in a way that it provides the relevant context to research questions, while discreetly shaping the way towards possible solutions.

## 2.1 Compliance

According to Lu et al. (2008), compliance means ensuring that business processes, operations and practice are in accordance with a prescribed set of norms.

The design of business processes has historically been driven by business objectives, but the importance of ensuring the compliance with regulations and standards has increased today and has become an important change driver. The authors further advocate that compliance should be implemented from an early stage of the design, in order to minimize the risks of compliance violations and penalties (Lu et al., 2008).

Maxwell and Anton (2010), state that the financial cost of non-compliance, together with the cost of lost reputation and brand damage, makes compliance a critical requirement in software systems (Maxwell and Anton, 2010).

In order to explain the compliance and its problems in relation to Data Lake concept and General Data Protection Regulation, one must first understand Big Data.

## 2.2 Big Data

There are many definitions of Big Data in circulation. Ohlhorst (2013) states that Big Data is often described as extremely large data sets that have grown beyond the ability of traditional data processing tools to manage and analyze them. However, the definition of the term Big Data is not limited only to nature of data sets; it also defines a situation in which data sets have grown to such enormous sizes that conventional information technologies can no longer effectively handle the size, scale or growth of the data set. In Big Data situation, the data set has grown so large that it is difficult to manage and extract value out of it. By managing Ohlhorst (2013) means acquisition, storage, searching, sharing, analytics, and visualization of data.

Ohlhorst (2013) further explains that the concept of Big Data has evolved to include not only the size of the data set or the situation in which the growth occurs but also the processes involved in leveraging the data. The concept has even become synonymous with other business concepts, such as business intelligence, analytics, and data mining.

Oracle (2015) describes Big Data as specific strategies to economically ingest, store, manage, analyze and share data. Special characteristics of this data have led to development of a new class of technologies and tools for accomplishing these operations. Zikopoulos et al. (2012) defines Big Data by using four Vs; volume, variety, velocity, and veracity. Definitions vary between different authors, but at least three of the Vs; volume, variety and velocity remain consistent. Oracle (2015), for example, uses value instead for veracity as the fourth V. The three Vs of Big Data, as shown on Figure 1, are:

- Volume, which stands for ever-growing size of the data sets. Zikopoulos et al. (2013) estimates the yearly growth rate of world’s digital data to be around 80%, which would mount up to 35ZB (35 trillion gigabytes) under the next few years. Ohlhorst (2013) states that enterprises easily amass terabytes and even petabytes of information. Oracle (2015), states that these high volumes consist of low-density data, or, in another words, data of yet unknown value such as feeds from social networks, clicks on a web pages, network traffic, captured data from sensors, and many more. According to Oracle, the task of Big Data is to convert all this low-density data into high-density data that has a defined value.
- Variety, which means that Big Data comes in very different data types. Zikopoulos et al. (2012) explains this variety by the need to try to capture all of the data that pertains to our decision-making process. Zikopoulos et al. further elaborates that unstructured data, such as audio files for example, can provide insights that were unimaginable by usage of traditional, structured data approach. Ohlhorst (2013) states that Big Data extends over structured data to include unstructured data of all varieties, while Oracle (2015) warns that unstructured and semi-structured data types, such as text, audio and video, require additional processing to both derive meaning and the metadata to describe it.
- Velocity, which is defined by Zikopoulos et al. (2012) as the rate at which data arrives at the enterprise and is processed. The value of Velocity is recognized by Zikopoulos et al. as ability to swiftly understand and respond to data signals, even in real-time. Ohlhorst (2013) states that maximizing value to the business dictates that Big Data must be used as it is streaming into the enterprise, but also that the data must still be available from the archival sources as well. Oracle (2015) observes that certain Internet of Things applications require real-time evaluation and action on Big Data. To support this observation, Oracle uses example of mobile application experiences that show large user populations, increased network traffic and the expectation for immediate response.

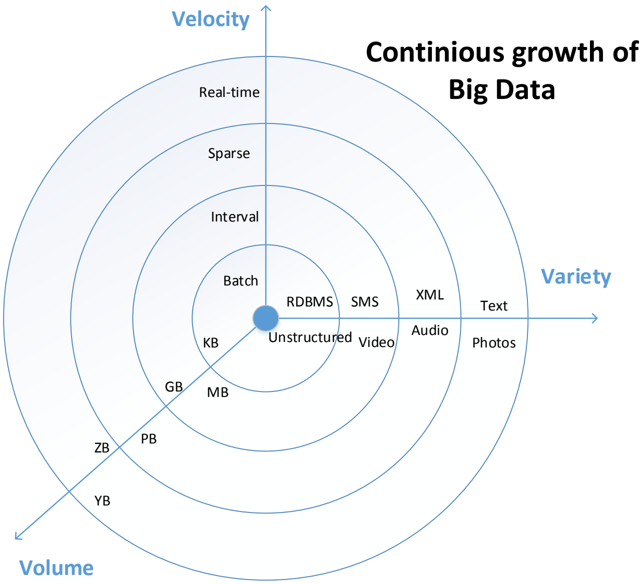


Figure 1, Big Data growth

Jacobs (2009), explains why distributed computing has become a dominant strategy for tackling Big Data. The reasons are price and scalability as it is more cost-effective to purchase and operate mainstream computer hardware that is both cheap and infinitely replicable than to build a specialized single system of great computing and storage capacity. Chaudhuri et al., (2011) identifies distributed systems using Map-Reduce Paradigm as very attractive for Big Data, as such platforms have the ability to support analytics on unstructured data in a scalable manner. The usage of such systems is growing rapidly, propelled by the availability of the open source Hadoop ecosystem.

Sawant and Shah (2013) recognize multiple challenges that follow Big Data. One of the very basic challenges is to understand and prioritize the usable data from garbage. The authors estimate that ninety percent of all the data is noise, which makes classifying and filtering a difficult task. Another challenge that is identified is that while searching for inexpensive methods of analysis, organizations have to compromise and balance against the confidentiality requirements of the data. Furthermore, usage of cloud computing and virtualization means hosting big data solutions outside the enterprise, which could become problematic out of confidentiality requirements perspective (Sawant and Shah, 2013).

### 2.2.1 Big Data and Compliance

Ohlhorst (2013) identifies compliance with data protection laws and regulations as a significant issue in Big Data environment. As a foundation for this statement, Ohlhorst claims that the current trend seems to be that businesses jump into Big Data, while forgetting about the specific pieces of information that may be mixed into their large data stores, which would expose them to compliance issues (Ohlhorst, 2013).

Kuner et al. (2012) share Ohlhorst's concern, stating that Big Data poses enormous challenges for data protection and concluding that there is little evidence that data protection can keep up with the pace of Big Data.

New data sources, such as mobile devices and social media applications, are generating massive amounts of unstructured data, giving a perfect example of a situation that causes exposure to compliance issues. According to Ohlhorst (2013), there are four important goals that an enterprise needs to reach in order to keep Big Data secure and in compliance:

1. Control access by process, not job function.
2. Secure the data at rest. This goal means that in practice, all Big Data, especially sensitive information, should be encrypted regardless of where it is stored.
3. Protect the cryptographic keys and store them separately from the data. The best practice in regard to regulatory compliance is storing the keys separately on a hardened server.
4. Create trusted applications and stacks to protect data from rogue users. This goal means that it is not enough to protect data with access controls. Access and configuration of the access controls themselves must be protected as well.

As the first security principle above all others, Ohlhorst (2013) identifies the need of knowing where the data resides (Ohlhorst, 2013).

Ohlhorst (2013) further claims that most of the Big Data environments in use today, including Hadoop, Cassandra and MongoDB, do not provide the necessary functionality for securing and managing confidential data.

The following chapter (see chapter 2.3 Data Lake) describes one such environment.

## 2.3 Data Lake

CITO Research (2014) describes the concept of Data Lake as closely tied to Apache Hadoop and its ecosystem of open source projects. CITO Research attributes the concept's popularity to its ability to provide a cost-effective and technologically feasible way to meet big data challenges. The concept supports the following capabilities:

- Capturing and storing raw data at scale for a low cost.
- Storing many types of data in the same repository.
- Performing transformations on the data.
- Definition of data structure at the time it is used (schema on read).
- Performing new types of data processing.
- Performing single subject analytics based on very specific use cases.

According to CITO Research (2014), the first usage of Data Lake concept was handling web data at Google, Yahoo, and other web-scale companies. Usage spectrum henceforth expanded to handle clickstream data, server logs, social media, geolocations, and machine and sensor data.

Stein and Morrison (2014) identify Data Lakes as an emerging approach to cloud-based Big Data that can be 10 to 100 times less expensive to deploy than conventional data warehousing. As an example of successful Data Lake implementation, the authors name UC Irvine Medical Center which used Data Lake based on Hadoop architecture to solve the problem of managing millions of records for more than a million patients, including radiology images and other semi-structured reports, unstructured physicians' notes, and volumes of spreadsheet data. According to Stein and Morrison (2014), there are four criteria that define Data Lakes:

1. Size and low cost; With Data lakes based on Hadoop, petabyte-scale data volumes are neither expensive nor complicated to build and maintain.
2. Fidelity; Because Data lakes based on Hadoop preserve data in its original form and capture changes to data and contextual semantics throughout the data lifecycle, compliance and internal audit is easier to maintain even after transformations, aggregations, and updates.
3. Ease of accessibility; In Data Lake, data in its original form is available throughout the enterprise, which eliminates internal political or technical barriers to increased data sharing.
4. Late binding; Structuring is task oriented (schema on Read) and does not require up-front data models (schema on Write).

(Stein and Morrison, 2014)

Stein and Morrison (2014) acknowledge that there is a risk of Data Lake initiatives producing Big Data graveyards. This happens when data is just dumped into the lake without keeping track of what's there. Preventing Data Lakes to become Big Data graveyards is achieved by means of creating, enriching, and managing semantic metadata incrementally (Stein and Morrison, 2014).

Capgemini (2013) describes the Pivotal Business Data Lake as a new approach that provides data to all constituents of the enterprise, using tools that can both be used to create a new Business Data Lake and to extend the life of existing Enterprise Data Warehouse (EDW in further text) solutions. Furthermore, Capgemini argues that this approach solves the issue of individual versus enterprise-wide views because individual business units can each get the local views they require, while there also exists a global view to meet enterprise-wide needs. This is achieved by using an integrated operational reporting platform across the entire enterprise.

According to Capgemini (2013), the Pivotal Business Data Lake solves the challenges that Big Data poses to EDW:

1. Reconciling conflicting data needs across the enterprise. EDW implementations create a single consistent view of information across the enterprise, while individual business units often need a local view that may require additional data elements specific to the business unit, which may not be relevant from a corporate/global perspective and therefore absent. Business Data Lake approach provides both global and local views of the same data.
2. Providing real-time access. Because EDWs are usually segregated from transactional and operational systems, they suffer from an inherent delay in information availability and data freshness that presents difficulties in handling real-time information. Business Data Lake approach solves this challenge through the use of performance-enhancing techniques like in-memory storage.
3. Assembling data from multiple sources causes a lot of time being spent on getting this data to a wide range of different analytical platforms. Business Data Lake approach provides a single analytic environment where data from multiple systems is ingested.
4. Supporting ad hoc analysis is something operational systems are not typically capable of without an adverse effect on performance. By providing parallelism, Business Data Lake's architecture overcomes these constraints and makes it possible to run ad hoc analysis as needed.

### 2.3.1 Data Lake and Compliance

According to EMC (2014), Hadoop-based Data Lakes can become a highly complex undertaking with a negative effect on security risks and operating expenses. In regard to meeting the compliance regulations, Hadoop lacks the capability to protect the growing amounts of data that it stores (EMC, 2014).

Harper (2014) recognizes the benefits of Data Lakes, such as cost reduction, expedience and eliminating the need for independent data marts, but at the same time Harper identifies several aspects of Data Lakes' effects on Big Data governance:

- Data Science, because while delivering access to data throughout an enterprise, the responsibilities of trained Data Scientists for managing data effectively fall to untrained end users.
- Metadata, because initial data attributes of semi-structured and unstructured data have yet to be determined.

- Semantics, because without a centralized instance that initially parses through the data to provision semantic consistency, user autonomy can result in disparate semantics.
- Risk/Security, because data collection without consistent definitions, Metadata, and semantics becomes subject to regulatory issues and security concerns. One of the Data Lake's unique selling points is enterprise-wide access to data that is largely undefined, which complicates regulating access to data by data type, use, department and other factors.
- IT, because on account of reducing architecture and simplifying the process of ingesting and storing data and Data Lakes can be viewed as a step back in technology that will affect performance and data quality in the long run.

Harper's (2014) conclusion after evaluating the effects of Data Lakes on Big Data governance is that the future of Data Lakes is dependent on their capacity to reinforce governance and address inherent security issues. Not solving these issues has the negative ramification of leading to the breach of regulatory compliance (Harper, 2014).

White and Heudecker (2014) are even harder in criticism of Data Lake concept, arguing that Data Lake concept per definition ignores how or why data is used, governed, defined and secured.

## 2.4 Failure of personal data protection laws previous to GDPR

Back in 2010, Paul Ohm wrote a very thorough article explaining the reasons behind the failure of implementations of privacy laws to protect personal data. According to Ohm (2010), the predominant reason that both the federal privacy statutes in the US and the Data Protection Directive in the EU failed to achieve their purpose is a wrongful assumption that anonymization of data protects privacy while keeping the data useful for data analysts. Data Protection Directive of 1995 is the current regulation for processing of personal data inside EU, which is to be superseded by GDPR.

Ohm (2010) uses three case studies to prove that advances in data analytics, namely reidentification science, thwart successful implementation of nearly every privacy law and regulation that depends on anonymization of data. The three studies that Ohm (2010) describes in his article are those of America Online, the state of Massachusetts, and Netflix. Each of these studies shows the ultimate failure of data anonymization and the concept of removing Personal Identifiable Information.

Anonymization is the process of manipulating information in a database in order to make it difficult to identify data subjects. Information that has to be manipulated in order to achieve anonymization is called Personal Identifiable Information (PII in further text) and what is considered as PII is accurately defined by a privacy law. One simple example of anonymization is a hospital data administrator suppressing the names of patients before sharing prescription data. Reidentification is the opposite process where anonymized data is linked with outside information, in order to discover the true identity of the data subjects (Ohm, 2010). The problem Ohm identifies with anonymization is that it relies on a balance between privacy and usefulness of data. Each removal of information from a dataset limits that dataset's usefulness for analytical purposes. Ohm argues very persuasively that advances in reidentification techniques has tipped the balance irrevocably and that: ”Data can be either useful or perfectly anonymous but never both.” (Ohm, 2010, p.1704).

Ohm considers European Data Protection Directive to impose a very strict set of requirements for notice, consent, disclosure and accountability on data administrators in order to force them into data anonymization. The administrators had a choice: “Anonymize your data to escape these burdens or keep your data identifiable and comply.” (Ohm, 2010, p.1763). Because of advances in reidentification, Ohm argues that this choice is non-existent on account that: “European privacy regulator can reasonably argue that any database containing facts (no matter how well scrubbed) relating to people (no matter how indirectly) very likely now falls within the Directive.” (Ohm, 2010, p.1763).

For that reason, Ohm predicted that European Union should reconsider lowering the floor of its comprehensive data-handling obligations while imposing heightened privacy regulations on particular sectors. This prediction did not come to pass as events took a very different turn in 2012 when a new regulation for processing of personal data inside EU was proposed in the form of GDPR.

## 2.5 General Data Protection Regulation

General Data Protection Regulation (referred to as GDPR in further text), is a new EU law that regulates processing of personal data. The draft of GDPR was proposed in 2012 and was adopted by European Parliament in 2014. Although the final technical details of the law are still being developed, the law’s principles are agreed on by both the EU Parliament and the EU Council and the law is expected to be finalized 2015/2016 and to come into force 2017/2018 (European Commission, 2014).

According to Triger (2014), GDPR is designed to unify and simplify data protection across the 28 member countries of the European Union (EU). It supersedes the Data Protection Directive of 1995 and any national legislation inside the EU. GDPR draft broadens the definition of personal data to any information relating to an individual, such as name, photo, an email address, bank details, postings on social networking websites, medical information or a computer's IP address. Triger (2014) states that the GDPR will impact all organizations, even non-European companies that operate in the EU, reflecting the fact that in the age of globalization, business has become borderless. As a consequence of the coming regulation, Triger (2014) argues that organizations will need to consider if and how they change the way they collect, process and store data. This interpretation of GDPR is confirmed by European Commission (2015), which argues for the need of same rules for all companies, regardless of their establishment. The Commission describes the current situation as uneven, because European based companies have to adhere to stricter standards than companies established outside the EU but conduct their business on EU market, thanks to Internet’s global nature (European Commission, 2015).

Villion (2012) lists the top five things businesses need to do to get ready for GDPR:

1. Organizations with more than 250 employees will need a data protection officer to act as the focal point for all data protection activities.
2. Information asset register must clearly identify what data is held, where, how and why.
3. Privacy policies must be written in plain English.
4. Processes and procedures will have to be implemented in order to handle data subject and data deletion requests.

5. Because of the extreme fines threatened for a serious breach, technical and procedural controls around data access would have to be reviewed.

According to Davis, M.'s (2014) summary of GDPR proposal that was adopted by EU Parliament 2014/03/12, all organizations processing the data concerning EU citizens, regardless of where the data is stored, would have to implement the following requirements:

1. Collect explicit consent to collect data from data subjects and provide the option of withdrawing that consent.
2. Facilitate "Right to erasure" – provide a service of deleting all data concerning the data subject upon request, unless there is a legitimate reason to deny said request.
3. Establish a clear privacy policy that will be provided to all data subjects.
4. Upon request, provide data subjects with a copy of their personal data in a format that can be transmitted electronically to another system.
5. Undertake annual risk management/analysis.
6. Decide on which is to be Single Data Protection Authority (DPA) for the organization.
7. Appoint a Data Controller inside the organization that would be responsible for compliance of all data processing operations with GDPR.
8. Appoint a lead authority Data Controller that would be responsible for all data processing operations across Europe.
9. Organizations processing data of more than 5000 data subjects would have to appoint a Data Protection Officer, responsible for organization's compliance with GDPR.

The proposition also includes joint liability of Data Controller and Data Processor (the cloud provider) for any breach (Davis, M., 2014).

Davis, M.'s interpretation of GDPR is in line with European Commission's factsheet that was released in 2015. Benefits that GDPR brings for citizens, according to the European Commission (2015) are:

- A right to be forgotten, which empowers individuals to request deletion of their personal data from the data processor.
- Easier access to your own data, which enables individuals to get a copy of their data in a portable format.
- Allows the citizens to decide how their data is used, by forbidding the processing of data without the explicit consent of data subject.
- The right to know when your data has been hacked, which forces companies and organizations to notify the national supervisory authority of serious data breaches as soon as possible.
- Data protection first, not an afterthought, which means that data protection safeguards should be built into products and services from the earliest stage of development, and that default privacy settings should be privacy-friendly.

(European Commission, 2015)

The penalties proposed for failing to enforce GDPR go up to 5% of global turnover of the whole organization, up to the maximum of 100000000 (one hundred million) EUR (European Parliament, 2014).

## 2.6 Metadata Repository

As already mentioned in chapter 2.2.1, Ohlhorst (2013) identifies the need of knowing where the data reside as the most important security principal in Big Data environment. Ohlhorst further states that metadata can be used to bring structure to unstructured data (Ohlhorst, 2013).

Brewer (2015) explains how metadata can be used to catalog the data on a Big Data platform, by using a metadata repository. Metadata repository is defined as a cataloging system that enables searches for data and access to it, on the base of metadata that is contained and managed in the repository (Brewer, 2015).

By its given definition, metadata repository provides the connection between data and the metadata that is describing it. Therefore the repository can be seen as a catalogue of connected information, providing the connection between the data's description and its location. The following chapter will provide the theory needed for deciding on an optimal solution for storage and management of connected data.

## 2.7 Storage and Management of Connected Data

Robinson et al. (2013) provides a comparison between three different approaches for storage and management of connected data:

1. Relational Databases, which are found to be a less than optimal choice because the rise in connectedness translates in the relational world into increased joins, causing increase in query cost and decrease in performance.
2. NOSQL Databases, which are also found to be a less than optimal choice because connections are established by one-way linked fields that point from one stored value to another. These links are powered by map-reduce, which introduces relative latency. Another problem is that links are done one-way, making a query of determining all data that is linked with a specific data instance a potentially expensive operation.
3. Graph Databases, which are found to be an optimal solution for storage and management of connected data because of the reasons presented in the text below.

According to Robinson et al. (2013), graphs are extremely useful in understanding a wide diversity of datasets in fields such as science, government, and business. The reason for this is that the real world does not follow the forms-based model behind the relational database; it is rich, interrelated and inconsistent in being uniform and rule-bound. Advantages of a graph database can be summarized in:

- Performance, when dealing with connected data in comparison to relational and NOSQL databases. Due to the fact that queries are localized to a portion of the graph, graph database performance tends to remain relatively constant, even as the dataset grows.
- Flexibility, which enables connecting data as the domain dictates, allowing structure and schema to emerge with our growing understanding of the problem space, rather than being imposed upfront. New relationships, nodes, and subgraphs can be added to an existing structure without disturbing existing queries and application functionality.

- Agility, which enables evolving our data model in step with the rest of our application, in alignment with modern incremental and iterative software delivery practices. The schema-free nature of the graph data model, combined with testable nature of a graph database's API and query language, makes it possible to evolve an application in a controlled manner.

According to Robinson et al. (2013), the advantage of graphs over other data modeling techniques is the close affinity between the logical and physical models. Robinson et al. argue that relational data management techniques introduce semantic dissonance between our conceptualization of the world and the database's instantiation of that model, which is considerably less apparent with graph databases. The graph model consists of nodes, relationships and properties:

- Nodes can be viewed as documents that store properties in the form of arbitrary key-value pairs. The keys are strings and the values are arbitrary data types.
- Relationships are connections between nodes. They always have a direction, a label, a start node and an end node. Relationships can also have properties, which provide additional metadata.

(Robinson et al., 2013)

## 2.8 Design Research

Maier (2010) describes Design Research, shown as Figure 2, as an integral part of the user–centered design process that is iterative in nature. At the start of the process, solutions are proposed based on observable phenomena related to the problem space. Then, a design solution is agreed upon and then prototyped in order to be tested against its target audience in the final step of the process. The process is repeated until satisfactory design is achieved. The purpose of each iteration is to add new context and insight to the design process.

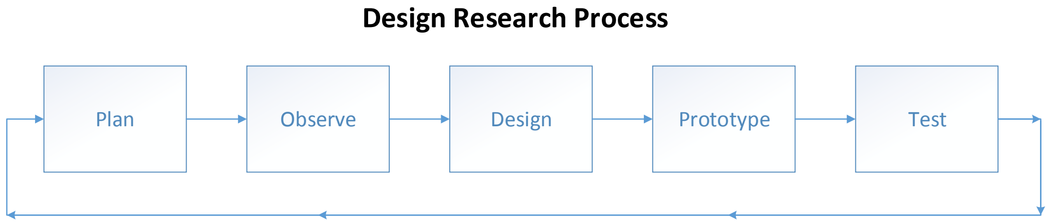


Figure 2, Design Research process (Meier, 2010)

According to Hevner et al. (2004), Design Research should result in an IT artefact that is created to address a specific, significant and relevant business problem. Usefulness, quality, and efficiency of a designed artefact must be determined through evaluation. Hevner et al. further postulate that Design Research must be presented in such a way that it's understandable to both technology-oriented staff as well as a management-oriented audience.

## 2.9 Enterprise Architecture Method

Lankhorst et al. (2013) defines an architecture method as a structured collection of techniques and process steps for creating and maintaining enterprise architecture. Lankhorst et al. further elaborate that methods typically specify phases and deliverables of architecture's life cycle. They also list currently used architecture development methods as:

- Rational Unified Process (RUP), with its extension towards enterprise architecture – Enterprise Unified Process. According to Lankhorst et al. (2013), this method provides an iterative and incremental alternative to the classical waterfall process.
- UN/CEFACT Modelling Methodology (UMM), which according to Lankhorst et al. (2013) is limited to a scope of business operations while omitting technology-specific aspects.
- TOGAF Architecture Development Method, which according to Lankhorst et al. (2013), provides a framework and development method for enterprise architectures, facilitating a detailed and well-described phases.
- Federal Enterprise Architecture Framework (FEAF), which according to Lankhorst et al. (2013) is designed for developing enterprise architectures specifically for governmental organizations.

Besides the above list, Lankhorst et al. (2013) notes that there are other, proprietary architecture development methods in use with various consulting companies.

### 2.9.1 TOGAF Architecture Development Method

The Architecture Development Method that will be used for development of Enterprise Architecture in this thesis is The Open Group Architecture Framework (TOGAF), Version 9.1. According to Lankhorst et al. (2013), TOGAF originated as a generic framework for development of technical architectures, but evolved into an Enterprise Architecture framework and method. It is specifically designed to address enterprise's business and IT needs by providing:

- A set of architecture views (business, data, applications, technology).
- Guidelines on tools for architecture development.
- A set of recommended deliverables.
- Linkages to practical case studies.
- A Method for managing requirements.

TOGAF ADM is an iterative process which is performed in phases, as shown on Figure 3:

- **Preliminary Phase**, which prepares the organization for a successful architecture project.
- **A: Architecture Vision**, which sets the scope, constraints and expectations for the project. The goal is to validate the business context and create the Statement of Architecture Work.
- **B: Business Architecture**, where baseline and target Business architectures are developed and gaps are analyzed.
- **C: Information System Architecture**, where baseline and target Information System architectures are developed and gaps are analyzed.
- **D: Technology Architecture**, where baseline and target Technology architectures are developed and gaps are analyzed.
- **E: Opportunities and Solutions**, where Major Implementation Projects are identified
- **F: Migration Planning**, where costs, benefits and risks are analyzed and Implementation Roadmap is created.
- **G: Implementation Governance**, where conformance between implementation projects and the architecture is ensured.

- **H: Architecture Change Management**, where it is ensured that the architecture responds to the needs of the enterprise as changes arise.
- **Requirements Management**, where validation against business requirements is done at every stage of the project.

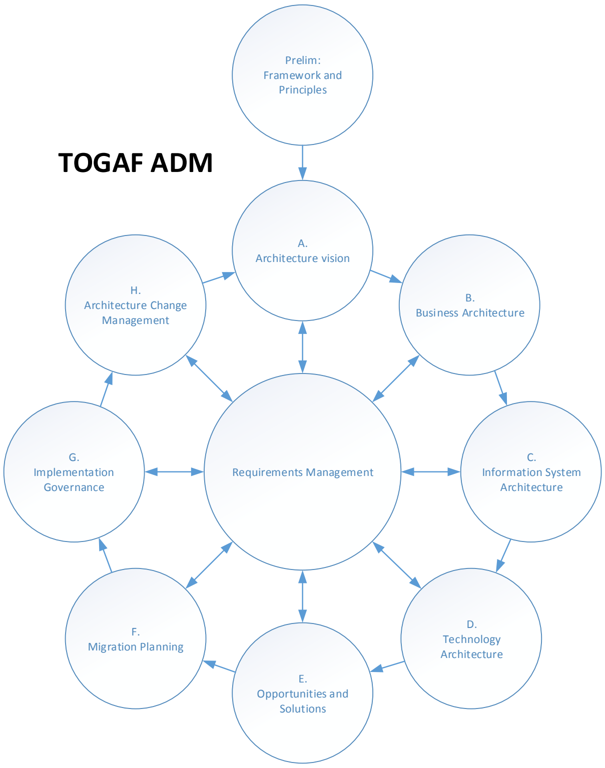

TOGAF ADM
Prelim: Framework and Principles
A. Architecture vision
B. Business Architecture
C. Information System Architecture
D. Technology Architecture
E. Opportunities and Solutions
F. Migration Planning
G. Implementation Governance
H. Architecture Change Management
Requirements Management

**Figure 3, TOGAF ADM (The Open Group, 2011)**

TOGAF ADM is a comprehensive general method that can be modified or extended to suit specific needs (The Open Group, 2011). Besides that TOGAF ADM is a method specifically designed for development of Enterprise Architectures, which is exactly what's being done under the course of this Thesis, and the fact that it is free to use, there is another argument for TOGAF ADM as method of choice. Capgemini, the main stakeholder of this project, has been involved in the development of TOGAF ADM and is a platinum member of The Open Group (The Open Group, 2015). A version of TOGAF ADM is being used internally by Capgemini for describing and developing EA, and Capgemini is also one of the providers of TOGAF ADM certifications. Therefore, the advantage of using TOGAF ADM lies also in the fact that the resulting design can be easily taken over and used by Capgemini for further development.

## 2.10 Open Kanban

Open Kanban project management method is Agile and Lean open source method for visualizing the workflow and incrementing the workload into reduced batch sizes that are easier to accomplish. This way, feedback can be received on every individual piece that's finished instead of waiting for feedback on the final product, which reduces the risk of wasting project resources that could lead to project failure. Kanban visualizes the workflow behind creating a product by using a Kanban board that enables a perspective on what stage the project is in by showing progress on each individual piece of the project (Hurtado, 2014).

# 3 Method and Result

The scientific method used in this thesis, Design Research, will be satisfied through the use of TOGAF ADM as the design method. Both methods are iterative and Design Research cycles can be mapped both to TOGAF ADM as a whole, as well as on its parts – phases or phase groups, in order to gain new knowledge through design.

The main reason for choosing TOGAF ADM as the design method is that it provides a complete framework for developing enterprise architectures. The framework is extremely detailed, which should contribute to reliability and validity of the results. Other reasons for this choice are that TOGAF ADM is an open source framework and that it is closely related and mainly compatible to the proprietary development method that is currently used by the project sponsor, which makes the results of this project easier to incorporate in future development work by the sponsoring organization.

Open Kanban will be used for planning and managing workload. Phases of TOGAF ADM will be incremented to the level of steps that are defined for each phase by the Open group (2011). Steps of TOGAF ADM phases will be managed as batch sizes, which shall enable a perspective on the status of the project at any given time.

As Fact-Finding Technique, background reading in combination with interviews was used according to good system development practices as described by Bennet et al. (2010).

## 3.1 Preliminary Phase

As a preliminary step to the ADM, a time schedule (see Appendix A) was created for interviews that are to be used as source of information for creating deliverables. Completed interviews' fields are colored green, while canceled interviews' fields are colored red in order to mark the status of the scheduled interview. An interview marked red represents an alert that new source of information must be found. The schedule is to be updated throughout the design method's lifecycle with status updates of the scheduled interviews, as well as with new interview opportunities if the need arises. The structure of interviews is semi-structured with open questions (also available in Appendix A), according to theory of Bennet et al. (2010), in a way that they cover specified topics, but allow the open discussion that can reveal new facts about the subject.

Conducting a kick-off interview with a member of Capgemini's staff 2015-02-15 marked the start of the design method's preliminary phase where questions such as where, what, why, who and how we do architecture will be answered. The conducted interview, coupled with supplied documentation about the system and available information about GDPR, gave valuable insight into key drivers, enterprise and requirements for architecture work. The Preliminary phase itself is conducted according to steps defined by TOGAF ADM 9.1 standard documentation supplied by The Open Group (2011).

### 3.1.1 Scope the Enterprise Organizations Impacted

As the first step of the preliminary phase, the scope of the impacted Enterprise Organizations is determined.

Core enterprise units that are affected and achieve most value from the work are those tasked with data management responsibilities that include governance, compliance and policy responsibilities. However, the subject of this work is a business-critical issue and as such it affects the whole enterprise.

Soft enterprise units, which will see change to their capability and work with core units but are otherwise not directly affected, are those tasked with data analysis because the structure of stored data would presumably change.

Extended enterprise units, which lie outside the scoped enterprise but may be affected in their own enterprise architecture are those with responsibilities over data sources because certain GDPR requirements, such as that of user's consent for data processing, would have to be implemented at the source where the data is being collected. As data sources can lie outside the scoped enterprise, they can be described as extended enterprise units.

No defined communities of stakeholders that would be affected by this work are identified at this point.

Governance involved that is of interest to this work is compliance. Legal frameworks that are found to be implemented at this point are European Data Protection Directive and a variety of USA laws that regulate data protection. It is important to note that these legal frameworks are not active in Business Data Lake enterprise architecture per default, but the EA fully supports these frameworks and Business Data Lake users are not constrained in any way by the EA in relation to compliance with above-mentioned legal frameworks.

It is important to note that the scope of this work is limited to Enterprise Architecture model, no deployed instance of Business Data Lake Enterprise Architecture falls inside the scope of this work.

### 3.1.2 Confirm Governance and Support Frameworks

Bringing produced architectural material (standards, guidelines, models, compliance reports, etc.) under governance is out of the scope of this work. Planning and managing the change of current enterprise governance and support models is a process that could really be started only after validation of produced artefacts. However, considering that compliance is a governance process that is already strongly supported by the current architecture framework, it would be logical to assume that bringing the new EA under the existing governance would not present an issue. That being said, even at this preliminary phase it is reasonable to assume that some GDPR requirements, such as those of reporting to Data Protection Authority and establishing Data Protection Officer, would mean changes to architecture governance organization and guidelines. It is clear that compliance as governance process is the main driver of the change and thus has ownership of the produced architectural artifacts.

Any potential impacts that could occur under organizational change are overshadowed by the fact that establishing compliance with GDPR is a business-critical issue.

### 3.1.3 Define and Establish Enterprise Architecture Team and Organization

Existing enterprise and business capability of the Business Data Lake are that of an established solution for managing Big Data and providing actionable insights throughout an enterprise, both from real-time data streams and from data at rest. According to Capgemini (2013), Business Data Lake has been deployed in large enterprises around the globe. In relation to compliance, there have not been any reported issues. However, it must be noted that out-of-the-box, Business Data Lake only provides the necessary tools and infrastructure to satisfy the governance requirements that can be different from each user case.

Architecture/business change maturity assessment is achieved through the kick-off interview with Capgemini's representative and through Business Data Lake Documentation that is summarized in the Appendix B. It is concluded that even though Business Data Lake provides certain flexibility in creating the governance to fit the specific user case, new GDPR requirements can prove to be impossible to address with current architecture and therefore a change might be needed.

As this project is performed as a part of a Bachelor Thesis at Luleå University of Technology, the work will be done individually by the Thesis' author. If the results of the Thesis are accepted, a proper Enterprise Architecture Team will be assembled to further the design. For the same reason, the financial budget assigned for this project at this point is non-existent.

Determined constraints on enterprise architecture work are:

1. Because the Enterprise Architecture of Business Data Lake allows the policies to be formulated according to a specific user case, the policies formulation will not be the focus of this thesis. Instead, the focus will lie on EA itself and eventual architectural changes that could prove necessary to enable implementation of new GDPR requirements, including policies.

Planning and managing the change of current enterprise governance and support models is a process that could really be started only after validation of produced artefacts and is therefore out of scope of this work. Some considerations about some aspects of this process may, however, be given in order to help future development.

### 3.1.4 Identify and Establish Architecture Principles

According to Open Group (2011), architecture principles are meant to be general rules and guidelines that inform and support the way an organization sets towards fulfilling its mission. Principles are identified from performed interviews and available Data Lake and TOGAF documentation. The list of identified Architecture Principles is available in the Appendix C (see Tables C1 – C22), sorted as:

- Business Principles
- Data Principles
- Application Principles
- Technology Principles

### 3.1.5 Tailor TOGAF

Due to constraint no. 2 (see chapter 3.2.3), which is about the validation of the produced artifacts before delivering of identified Target Architecture to the enterprise, TOGAF ADM phases will be performed up to phase D: Technology Architecture, with Gap analysis as a final product of these phases. Another important reason for this tailoring is that the design work being done is relevant to Enterprise Architecture model, not to deployed Enterprise Architecture.

### 3.1.6 Implement Architecture Tools

As a modelling language, ArchiMate 2.1 will be used. The motivation for this choice is that ArchiMate is fully aligned with TOGAF ADM 9.1 as its structure corresponds with three levels of architecture addressed in ADM phases B, C and D - Business, Information Systems and Technology Architecture. According to Open Group (2015), the language enables the creation of fully integrated models of the organization's enterprise architecture and is fully aligned with the TOGAF standard (The Open Group, 2015).

Standard office productivity applications will be used for creating and communicating deliverables that support the design. Visual Understanding Environment mind mapping software will be used for creating conceptual architecture vision.

### 3.1.7 Preliminary Phase deliverables

Most of the outputs of the Preliminary Phase of TOGAF ADM 9.1 are presented in the current section 3.2. In addition, Request for Architecture Work document, that provides valuable information about project's parameters, was created on the base of kick-off interview and its copy is attached to this report as Appendix D.

## 3.2 Architecture Vision

Phase A of the TOGAF ADM started with creation and agreement on Request for Architecture Work document (see Appendix D). According to The Open Group (2011), the purpose of this phase is to develop a high-level aspirational vision of the capabilities and business value that the proposed enterprise architecture will result in. This vision will be presented as a first-cut, high-level description of the Baseline and Target Architectures.

Architecture vision phase is conducted according to steps defined by TOGAF ADM 9.1 standard documentation supplied by The Open Group (2011).

### 3.2.1 Establish the Architecture Project

This Architecture Project is a standalone project, conducted for a specific purpose. Because the project is done as a Bachelor Thesis by one person, there is no project group that needs to be managed and coordinated. Architecture activity will be planned and managed using Open Kanban project management method, which will visualize the workflow and increment the workload into reduced batch sizes that are easier to accomplish. Risk of project failure is reduced because feedback can be received on every individual piece that's finished instead of waiting for feedback on the final product that may be negative, making all invested time, work and resources a clear waste. Workflow visualization on a Kanban board will enable a perspective on the project's progress. The workload's increments are mapped into steps of TOGAF ADM phases, with phases themselves as milestones.

The endorsement of the project by corporate management has been communicated at this point through Capgemini's representative.

### 3.2.2 Identify Stakeholders, Concerns, and Business Requirements

Stakeholder Map is created (see Appendix E, Table E1) on the base of available documentation and conducted interviews. The focus of the Stakeholder Map is set on stakeholders of the compliance issue with GDPR. A less detailed list of stakeholders is presented below. The first iteration of the stakeholder map is based on the available interview – the kickoff interview and available documentation at the start of the first iteration of the Architecture Vision Phase. Along with iterations of ADM phases, stakeholder map can be expected to grow as new stakeholders and new concerns/requirements are discovered:

- Project sponsor – Capgemini as the vendor of Business Data Lake solution.
- Project Management Group – business unit inside the project sponsor that is assigned with managing this project.
- Users - organizations that use Data Lake based solutions to achieve their business goals and process personal data of EU citizens.
- Competition – competing vendors of other Big Data solutions that are interested in solving the same compliance issue with GDPR that Business Data Lake product has.
- Data Protection Authorities – government agencies assigned with enforcement of GDPR.
- Data custodians – persons responsible for data management, more precisely Data Protection Officers, according to GDPR.
- Software engineers and other technology experts – technology experts within the enterprise that is tasked with development and support of the technology behind the Business Data Lake: Enterprise Architects, Data Scientists, System and Database administrators, System and Software Developers, and many more.
- Data subjects – persons that the processed personal data refers to.

The Stakeholder Map lists the first draft of stakeholder's concerns and requirements, while it also lists the communication plan, estimated decision power and level of interest for each stakeholder. Furthermore, the map contains the listings of views and artefacts that are to be made available for respective stakeholders.

### 3.2.3 Confirm and Elaborate Business Goals, Business Drivers, and Constraints

In this step of the Architecture Vision phase, business goals, drivers and constraints of the organization that are listed in the Request for Architecture Work document are confirmed and elaborated (see Appendix F).

### 3.2.4 Evaluate Business Capabilities

In order to be able to pinpoint those business capabilities which are candidates for improvement through architecture design, the overall Business Data Lake model is broken down into separate business processes, as shown on Figure 4. As the basis for process identification and evaluation, a summary of information that was collected through documentation (see Appendix B) and interviews was used.

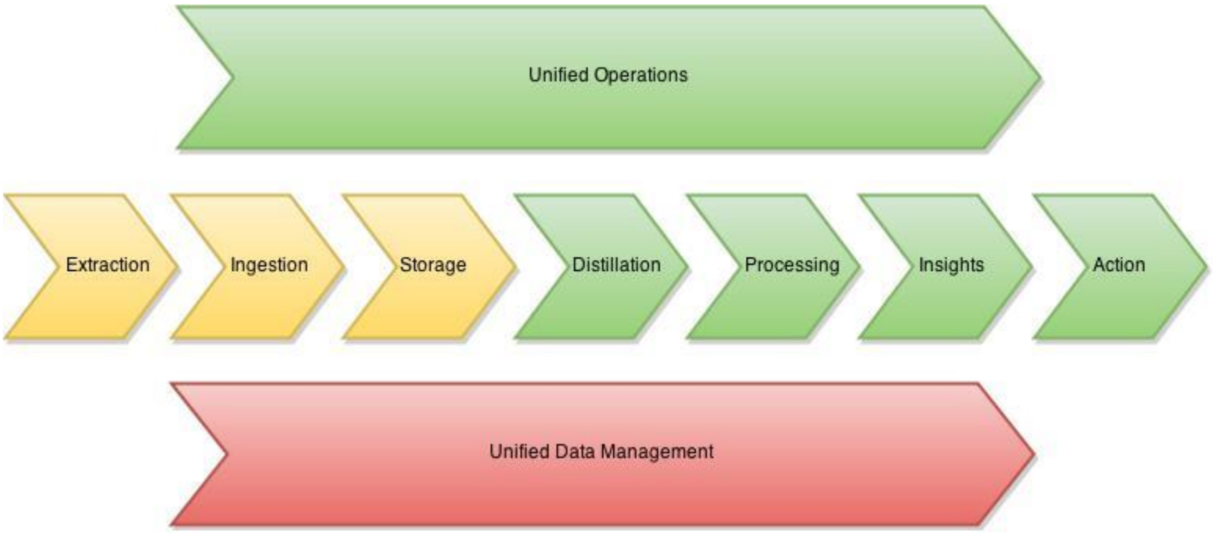


Figure 4, Business Capabilities

Business Data Lake is marketed as a universal solution for managing and deriving actionable information from Big Data. Its business capabilities have been proven through real life deployment. However, there is an evident room for improvement of capabilities in certain areas, especially in relation to compliance.

The **Extraction** process is technically outside the Data Lake system. It is a process of collection of data from it sources, before the data is ingested into a Data Lake. The sources of data come in great variety, which means that they can be from different systems, in different formats and different timeliness. The standard and descriptiveness of metadata differs as well. This consequently makes recognition and classification of incoming data a difficult task that can be reliably completed only through Distillation and Processing stages. As these processes are invoked only when needed to support the analysis of data, it leaves a possibility for data to be stored in a Data Lake without knowing what this data actually is. In relation to compliance with personal data regulations, this creates a space for unregulated, and therefore in collision with said regulations, collecting and storing of personal data.

**Ingestion** is the process where collected data is directed to its designated storage. It stands for capability to bring data from multiple sources across different timelines with varying quality of service. Ideally, this is the point where metadata repository should be populated in connection with the data being sent to storage. As this situation is today, this is not the case, as the metadata is only partially collected, through distillation and processing.

**Storage** is the process that stands for capability to store all forms of data in a cost-efficient manner. This is true of both unstructured and structured data of all types. A cost-efficient storing is achieved through scalability of Hadoop. This capability may be impaired due to new compliance requirements that are not compatible with the current way Hadoop Distributed File System is used, due to a technical constraint of Hadoop that causes immutability of data.

**Distillation** stands for capability to take the data from storage and convert it into structured data that can be processed easier. That can be described as an ETL operation during which the ingested data is converted into analytics-friendly format through transformation and aggregation. When needed, data is distilled from its raw form and given structure so that it can be processed and acted upon. Metadata repository is primarily filled through Distillation and through subsequent Processing.

**Processing** stands for capability to use analytical algorithms and user queries in order to analyze data and to generate structured data that can be used for further analysis. Analysis of data using Hadoop analytical interfaces such as Hive, HBase, Pig and MapReduce can be done in parallel with query processing through the ability to manage cross-platform workflows.

**Insights** stands for the capability to analyze the processed data and generate insights that support business decisions. Through the use of interactive analytical tools and dashboards, business users can join data across different data sets and draw insights. Triggering of external events can be done through real-time insights, which can be done automatically as insights are created in real-time as the data is being ingested.

**Action** stands for the capability to integrate insights with systems for making business decisions. This enables creation of data-driven applications where actions are performed automatically based on changes of business environment even as they happen, thanks to the real-time timeline that is consequently supported across all business processes from data extraction to action on collected data.

**Unified Data Management** stands for the capability to manage data throughout the data lifecycle in the Business Data Lake. It also enables definition of access policies. Finding the data is enabled through a central metadata repository. The data is then copied to a sandbox environment where it is distilled, processed and analyzed in order to provide actionable insights. Management of master data and reference data management services are capabilities that should exist but are currently not supported. This is a major problem because it hinders achieving compliance in the area of management of personal data.

**Unified Operations** stands for the capability to monitor, configure, and manage the whole platform of Business Data Lake from a single unified operating environment.

### 3.2.5 Assess Readiness for Business Transformation

According to The Open Group (2011), assessing the readiness of the organization to accept change of Enterprise Architecture and identifying acceptance issues is important for the success of ADM phases E and F. These phases are out of scope of this work and readiness assessment may well change by the time of eventual further development, but a current readiness assessment will be provided nevertheless in order to describe the circumstances of this project. Readiness is assessed on the base of available information about the project sponsor and on the base of project sponsor’s commitment to this project.

| Readiness Factor | Urgency | Readiness Status | Degree of Difficulty to Fix |
|---|---|---|---|
| 1. Vision | No | High | No action needed |
| 2. Desire/willingness/resolve | No | High | No action needed |
| 3. Need | No | High | No action needed |
| 4. Business case | Yes | Acceptable | Easy |
| 5. Funding | Yes | Low | Moderate |
| 6. Sponsorship and leadership | Yes | Acceptable | Moderate |
| 7. Governance | Yes | Low | Difficult |
| 8. Accountability | Yes | Low | Moderate |
| 9.Workload approach and execution model | Yes | Fair | Moderate |
| 10. IT capacity to execute | No | Good | Easy |
| 11. Departmental capacity to execute | No | Good | Easy |
| 12. Ability to implement and operate | No | Good | Easy |

Table 1, Readiness for Business Transformation

Table 1 shows the assessment of readiness factors for business transformation. If a readiness factor is assessed as urgent, it means that an action is needed to improve readiness status before a change can be started.

Readiness status can be described as:

- Low, which means that substantial work is needed before proceeding with change.
- Fair, which means that some work is needed before proceeding with change.
- Acceptable, which means that readiness issues exist, but are not critical.
- Good, which means that only minor readiness issues exist.
- High, which means that there are none readiness issues.

Degree of Difficulty to Fix can be described as: no action needed, easy, moderate or difficult.

### 3.2.6 Define Scope

According to The Open Group (2011), there are four dimensions that are typically used to define and limit the scope of architectural activity:

1. Breadth, which is, for this project, limited to those parts of the enterprise architecture where a need for change is identified. A high abstraction level of the whole architecture will be provided for the purpose of understanding the business process as a whole, while more detailed levels of architecture will be used to describe those parts of the enterprise architecture of relevance to compliance problem and the design of a solution. At this point, parts of enterprise architecture that will be described in greater detail are identified as those related to Ingestion, Storage and Unified Data Management business processes.
2. Depth, which will, for this project, be developed only for parts of the enterprise architecture of relevance to compliance problem and the design of a solution. For these parts, the level of depth will show the detail of a single data instance across all four architecture domains.
3. Time Period, which will, for this project, show a baseline instance of a current state of architecture and an instance of a target state of architecture.
4. Architecture Domains, which are planned to be addressed in full scope as enterprise architecture will be described in all four domains: Business, Data, Application and Technology domain.

Standardization of metadata at data sources is out of the scope of this work, as this is an issue that is beyond the control and responsibility of this project.

Design of usage policies is out of scope of this work, as this is dependent on a specific case of a deployed Business Data Lake.

Planning and managing the change of current enterprise governance and support models is out of scope of this work, because this process that could really be started only after validation of produced artefacts.

### 3.2.7 Confirm and Elaborate Architecture Principles, including Business Principles

Architecture Principles that were defined during the Preliminary phase of ADM and attached to this document (see Appendix C, Tables C1 – C22) are confirmed through email communication with Project sponsor on 2015-03-17.

### 3.2.8 Develop Architecture Vision

The Open Group (2011) defines Architecture Vision as a high-level view of the Baseline and Target Architectures, based on the stakeholder concerns, business capability, requirements, scope, constraints, and principles.

#### *3.2.8.1 Baseline Architecture Vision*

The first step in developing Architecture Vision is creating a high-level concept of a current architecture state – Baseline architecture. The concept presented on Figure 5 describes the major components and operation of Business Data Lake at the start of this project.

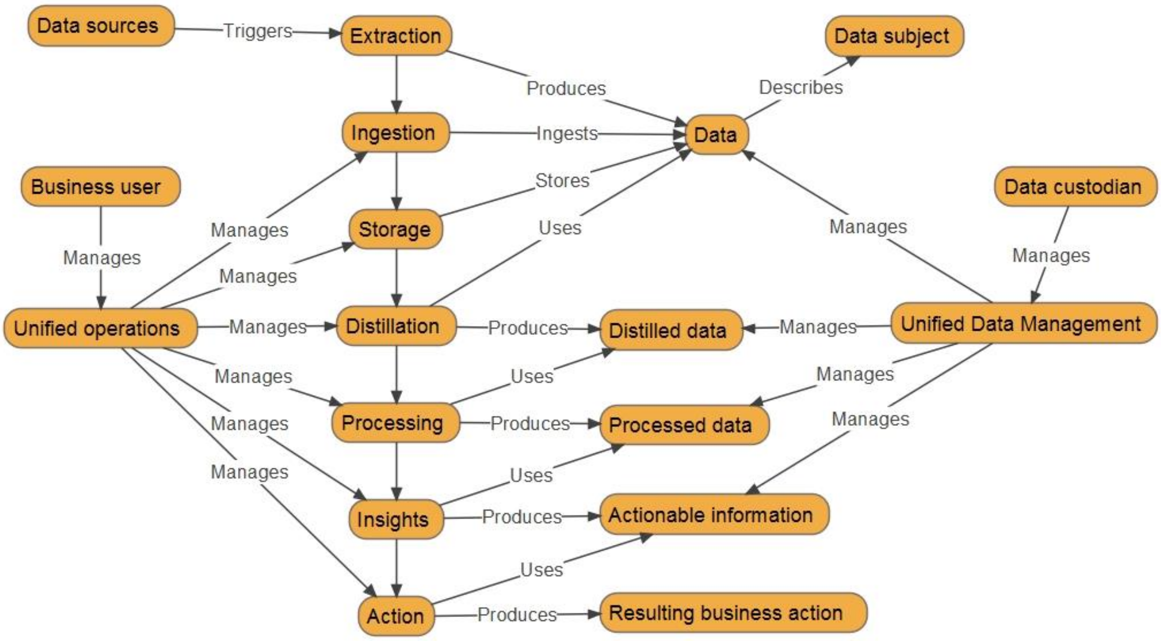


Figure 5, Baseline Architecture Vision

Business scenarios for the baseline architecture are described for each participating stakeholder:

### Business user

Business user has the capability to monitor, configure, and manage the whole platform of Business Data Lake by using a single unified operating environment. When data is extracted from data sources, it is ingested into Business Data Lake by rules that set by a Business user according to intended usage of data. For example, data streams that require real-time analysis will be ingested directly into in-memory database to be processed and analyzed accordingly and later released on disks to become data at rest. Other data may be ingested into Hadoop directly, to be processed and analyzed according the need of a business user. The user can review and set the rules for each process that the data goes through and can also access and work with results of these processes, limited only by the access rights the user possesses.

However, the environment and data access is strictly controlled. When a Business user requests access to the data, a sandbox working environment is created where only the data that the Business user has access rights for can be copied and where only those operations that the user has access to are allowed. The sandbox exists until the user is finished with his work, but the environment's existence is also time limited, after which the copied instance of the data, as well as the newly generated data is released back into the data pool.

### Data custodian

At the current state of architecture (Baseline), data custodian monitors, configures, and manages access rules to data, distilled data, processed data and actionable information through Unified Data Management environment. As the data is ingested and stored into the Business Data Lake, data custodian gathers existing metadata, sets the descriptive metadata where it is possible under the constraints of the ingestion process and populates the metadata repository that is used to locate the desired data on the system, according to desired criteria.

As the data is distilled and processed, new data and metadata describing the original data is generated through aggregation and analysis. It is the responsibility of data custodian to manage access rules and metadata even for this new data, as well as to update the metadata of the original data.

Data custodian is also responsible that the personal data is identified, used and managed according to relevant legislation and policies. This is however not a separate nor consistent process, as the personal data can be identified and located only through metadata repository where metadata for all types of data is stored in inconsistent manner.

### Data subject

At the current state of architecture (Baseline), data subject has data inside the Business Data Lake that can be classified as his personal data. The only interaction that data subject is a currently a part of is done outside the Business Data Lake system, at data sources/extraction, when he gives consent for his data to be used and processed by a second or third party.

The abovementioned business scenarios are developed further into use cases and functional requirements. Detailed description of these use cases is available in Appendix G (Figures G1 – G13, Tables G1 – G13), while a detailed description of requirements is available in Appendix H (Tables H1 - H38).

### *3.2.8.2 Target Architecture Vision*

The next step in developing Architecture Vision is creating a high-level concept of a desired architecture state – Target architecture, presented as Figure 6. As a resource in creating a vision of target architecture, updated business scenarios were used to discover use cases (Appendix G, Figures G14 – G22, Tables G14 – G22) and document business requirements (see Appendix H, Tables H38 – H49) that the target architecture aims to satisfy.

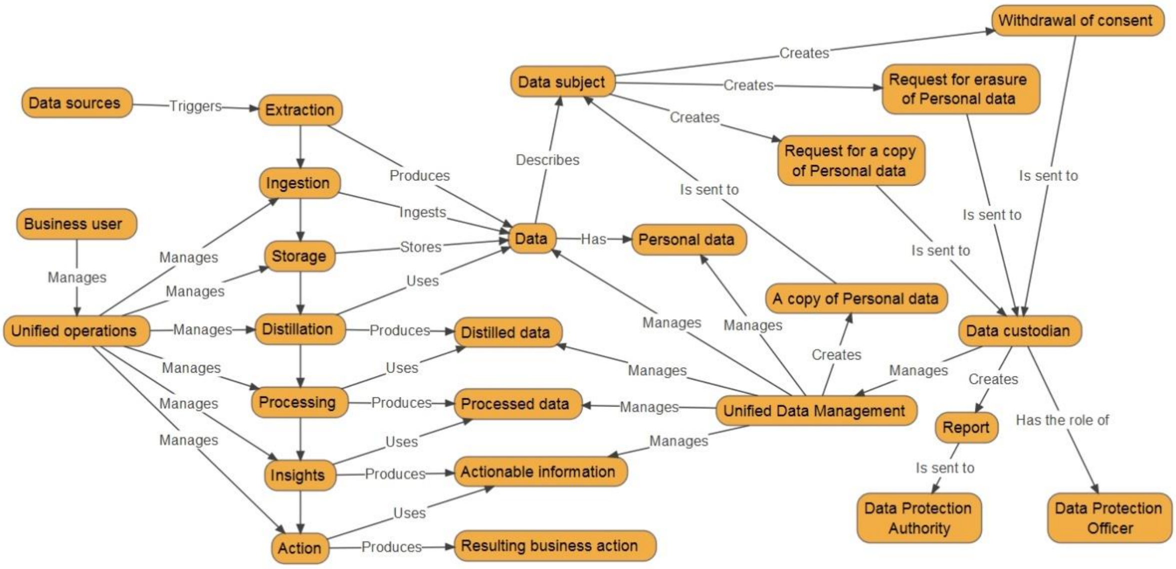


Figure 6, Target Architecture Vision

Business scenarios for the target architecture are described for each participating stakeholder:

**<u>Business user</u>**

Business scenario involving Business user remains the same as in Baseline Architecture Vision, as no legitimate business need exists to motivate the change.

**<u>Data custodian</u>**

Business scenario involving Data custodian has changed significantly in Target Architecture Vision. Core tasks inherited from Baseline Architecture Vision remain and are still performed as in Baseline Architecture, following the architecture principle that no architectural changes shall be made unless motivated by a legitimate business need. Management of personal data stands for the difference between the two architectures and the legitimate business need that caused this change is that the architecture must be improved in order to satisfy new compliance-related requirements.

Apart from usual duties, Target Architecture Vision tasks the Data custodian with assuming the role of Data Protection Officer. This role makes the custodian responsible for enforcement of compliance legislation across the organization.

The first new business scenario is that data custodian must respond to a Request for copy of personal data that is submitted by Data subject. Data custodian must find all personal data belonging to the Data subject, copy it into an interchangeable format and send it to the Data subject. In order to be able to achieve this task, a register of personal data must be strictly maintained. The reason for this requirement is that it's not optimal to perform a search through all the data in a Big Data environment, making a specialized register for personal data a more viable solution. Implementing the register as graph database would optimize the search even further.

Another scenario where a need for a personal data register is evident is when a Request for erasure of personal data is received by Data custodian. Again, the personal data must be located in order to be erased, which can be solved with previously mentioned specialized register for personal data. However, this business scenario is much more complicated than the first one. Inside the Business Data Lake, data is constantly processed, analyzed and aggregated into new data. These processes are described in Baseline Architecture Vision, inside the business scenario involving the Business user. New data that is developed from processed data can also be assumed to contain personal data of one or many data subjects. As data is copied and processed by Business users inside separate time- and access-limited environments called sandboxes, deleting instances of this data that remain in ordinary storage does not provide a guarantee for a Data subject that his personal data is fully erased. If the personal data is being processed while the erasure takes place, after the processing is done the data would be returned to storage, leaving the organization in clear violation of Data subject's Right to erasure compliance requirement.

The solution for the abovementioned scenario is keeping track of all instances of personal data, not just of the original instance that lies in storage. That way, an erasure would be completed after the data is released from sandboxes into storage, during which the new instances of data would be registered into personal data register and also erased if a relation to the Data subject who submitted the Request for erasure is established.

A third business scenario involving Data custodian would be responding to Withdrawal of consent request that is submitted by Data subject. It is still not entirely clear what consequences the consent withdrawal has on data processing. It will be assumed at this point that the consent would be revoked for all data that would be ingested after the arrival of consent's withdrawal and that all the data that was ingested before this point would still remain available for processing. Collecting of data takes place outside of the Business Data Lake system and is therefore out of the scope of Data custodian's duties. However, Data custodian can ensure that no data that are covered with consent withdrawal exist on the system.

Upon receiving Withdrawal of consent from a Data subject, it is Data custodian's duty to make sure that no new personal data connected to subject in question is ingested and stored from that point.

A forth business scenario involving Data custodian would be reporting to Data Protection Authority about serious security breaches in data management and compliance. If the security of personal data should be compromised or if the way that data is managed and processed breaches against the relevant legislation, Data custodian is obligated to report this to Data Protection Authority as soon as possible. This duty is even more emphasized if Data custodian has the role of Data Protection Officer, which is highly probable in a Big data environment such as data lakes.

**Data subject**

Precisely as in the current state of architecture (Baseline), data subject still has data inside the Business Data Lake that can be classified as his personal data. The consent for data to be collected and used is still given at data sources/extraction. In the Target Architecture Vision, Data subject has the option of withdrawing that consent, by a request made to Data custodian. Upon sending the Withdrawal of consent, no subsequent data should be collected and used from sender. It is a matter of some controversy if the consent can be revoked for the data that has been collected previous to receiving the request for consent's withdrawal. It will be assumed that this is not the case because Data subject can easily achieve such a result by pairing Request for erasure of personal data together with Withdrawal of consent.

Another business scenario revolves around sending Request for erasure of personal data by a Data subject. After the data subject makes such a request, it is his expectation that all personal data that relates to this data subject should be erased from Business Data Lake.

Third business scenario involving Data subject is sending a Request for copy of personal data. As a response, Data subject should receive the copy of all personal data relating to him. This copy should be in a format that can easily be transferred across systems.

The abovementioned business scenarios are developed further into use cases and functional requirements of Target Architecture Vision. Only those use cases and requirements that are changed from baseline or are completely new are listed under Target Architecture Vision subtitle. Detailed description of these use cases and requirements is available in Appendix H.

### 3.2.9 Define the Target Architecture Value Propositions and Key Performance Indicators

As the Target Architecture Vision shows, the solution will benefit the enterprise in that those compliance requirements that arise with the coming of GDPR are satisfied. More concretely, Right to be erased, Right to get a copy of your personal data and Right to withdraw consent are value propositions of the target architecture that would satisfy not only Data subject stakeholder group, which would be able to realize their newly given rights, but also Project sponsor, Users and Data Protection Authorities as an important step towards GDPR compliance.

Enabling consequent management of personal data and of Data subjects through a dedicated register would primarily be appreciated by Data custodian's stakeholder group, as they would be given the means of solving compliance requirements related to GDPR. Furthermore, business tasks involving personal data would be optimized, creating the direct value proposition for User's stakeholder group. Indirectly, this would create value even for Project sponsor, through making the Business Data Lake product more competitive.

Project sponsor's main value proposition of Target Architecture is that it addresses the compliance requirements of GDPR, which are business critical in nature. If these requirements are not addressed before the time GDPR is due for enforcement, the continuance of Business Data Lake as a product would be problematic.

The Key Performance Indicator of the Target Architecture would be achieving the compliance with GDPR. It is difficult to quantify this indicator in greater detail than a simple yes or no, for the reason that partial compliance does not qualify as a solution.

Another important performance indicator is overall performance of Business Data Lake in data processing. Although this indicator is difficult to quantify without doing performance measurement on a deployed Business Data Lake, the Target Architecture is required not to degrade this performance in any significant way. However, when it comes to processing of data, the Target Architecture offers an additional and quick way to find and select personal data, that in no way interferes with baseline functionality, as can be seen by comparing the two architecture visions. Therefore it can be said that Target Architecture Vision proposes the value of even faster data processing than the Baseline Architecture Vision, by providing a quicker way to filter out and select data based on a condition of data's relation to a Data Subject. The direct means of achieving this value proposition is through a dedicated registry of personal data.

Again, it is difficult to set a quantifiable Key Performance Indicator for improvement of overall performance of Business Data Lake. The reason for this is that quantification would require knowledge of the exact participation of data that is being processed and analyzed by relation to its Data subject in the overall data processing and analysis. It can however be said that the overall performance would be improved through implementing a dedicated personal data registry.

### 3.2.10 Identify the Business Transformation Risks and Mitigation Activities

Not achieving the target state would create business critical consequences for enterprises that use Business Data Lake. Threatened penalties that go up to 5% of global turnover for the whole enterprise or up to 100M euros are a strong deterrent as they effectively make non-compliant solutions for data processing unusable after the due date for enforcement of GDPR.

Identified initial business transformation risks and mitigation activities are classified according to TOGAF ADM documentation (The open Group, 2011) and are available as Table I1 under Appendix I.

### 3.2.11 Develop Statement of Architecture Work; Secure Approval

According to The Open Group (2011), Statement of Architecture Work document is created on the base of Architecture Vision Phase's deliverables. The statement should be communicated to Project sponsor's representative for approval, before the development can be continued.

Instead of producing a separate Statement of Architecture Work document with content that mirrors the deliverables of Preliminary Phase and Architecture Vision Phase, a copy of this report is sent to Project sponsor for review and adoption. The reasons for this decision are:

1. To save valuable time that can be better used for actual development in a time-limited project.
2. To make reading of this report a more enjoyable experience by limiting duplication of text.

Additional information that is specific for Statement of Architecture Work is presented in the text below.

The emphasis of this project is on proposing the desired target architecture. Baseline architecture shall therefore be described in a higher level of abstraction, while the target architecture shall be described in greater detail with additional views that support business objectives, with a concrete representation of solutions.

Business objectives of target architecture for this project are identified as satisfying the following compliance requirements of GDPR:

- Implement management of personal data
- Implement the Right to erasure
- Implement the Right to receive a copy of personal data
- Implement Withdrawal of consent

These compliance requirements are chosen as objectives because they implicate the need for a system change in personal data management. Furthermore, it is these requirements that are recognized as central points of GDPR and the most significant change in target architecture. GDPR brings other requirements as well, such as that of reporting security breaches to Data Protection Authority, but these are not identified as drivers of a significant change in architecture and the way personal data is managed as the listed requirements.

Other project parameters, such as stakeholders, timeline, roles, and work plan remain the same as in Request for Architecture Work document.

### 3.2.12 Architecture Vision Phase Outputs

Most of the outputs of the Architecture Vision Phase of TOGAF ADM 9.1 are presented in the current section 3.3. In addition, use cases' and requirements' details are attached to this report as Appendix G (see Figures G1 – G22 and Tables G1 – G22) and Appendix H (see Tables H1 – H49), respectively.

## 3.3 Business Architecture

Business Architecture phase of TOGAF ADM will result in design of the baseline architecture – the current state and the target architecture – the desired state. The phase will be concluded with Gap analysis, identifying gaps between the baseline and target architectures.

### 3.3.1 Select Reference Models, Viewpoints and Tools for Business Architecture

Baseline Architecture Vision will be used as a high-level reference for baseline of business architecture. Reference models described in Business Data Lake documentation will be used as a complementary information source in developing baseline architecture views. Relevant Business Architecture views are:

1. General view, which is a high abstraction of business processes of Business Data Lake that gives context to more specialized views of more detail.
2. Data custodian's view, for the reason that it describes the architecture area where most target requirements point to as an area where a change is needed.

3. Data subject’s view, because Data subject’s role changes from passive to active between baseline and target state, which prompts for a new business architecture to support this change.

Due to consideration towards readers the final version of this report will contain a detailed description of General view of Business Architecture, while only a gap analysis of other listed views will be included. Specialized views of Data custodian and Data subject are however available in great detail and can be provided upon request.

Tools and techniques identified as appropriate for capture, modelling and analysis are business scenarios, use cases and baseline requirements that were produced during the Architecture Vision phase, as well as business architecture principles that were identified during the preliminary phase of TOGAF ADM.

Actors identified for general view are Business user, Data custodian and Data subject.

Actors identified for Data custodian’s view are Data custodian and Data subject.

### 3.3.2 Develop Baseline Business Architecture Description

Baseline Business Architecture describes the current state of Business Architecture.

#### *3.3.2.1 General view*

General view, shown as Figure 7, provides a high level abstraction of Business Data Lake’s baseline business architecture. The inputs for diagrams’ design are business scenarios, use cases and baseline requirements identified during Preliminary Phase and Architecture Vision Phase of TOGAF ADM.

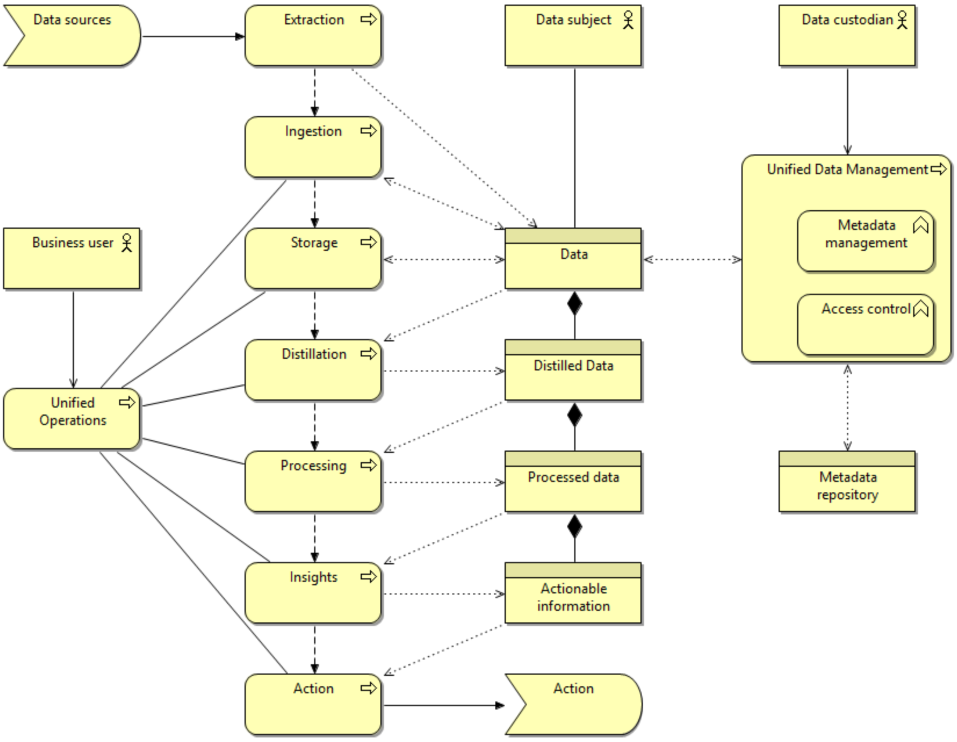


Figure 7, General view of business architecture baseline

**Process description:**

The process flow starts with the **Extraction** of data from various types of data sources (social networks, activity-generated data, sensory data, data warehouses, etc.). Extraction of data is actually outside the scope of Business Data Lake model as the extraction technique and its rules are dependent on the sources themselves.

**Ingestion** process directs the extracted data to its appropriate storage. The process is directed by business rules that are set by a Business user through Unified Operations process. The data is directed to a storage that best suits the business needs of processing and analysis that is planned for that particular data.

**Storage** process stores the data and makes it available for other processes. The way the data is stored is indirectly set by a Business user who, through Unified Operations, starts those business processes that use the data.

**Distillation** process distills the stored raw data by giving it more structure so that it can be selected and used for further processing. Business rules, that are set by a Business user Unified Operations, regulate the distillation parameters.

**Processing** applies analytical and machine-learning algorithms to selected data so that insights can be extracted from it. Processing parameters are regulated by a Business user through Unified operations.

**Insights** process enables joining of data across different data sets in order to draw insights through the use of interactive analytics tools and graphical dashboards. Data can be analyzed by using MADlib analytic libraries for performing mathematical, statistical, and machine learning analysis, as well as real-time insights that can be set to generate external events.

**Action** process enables both manual and automated making of business decisions based on analysis of data. Data-driven applications can be made by coupling insights with business decisioning systems.

**Unified Operations** is a process that enables the Business user to manage the whole data lake through a single unified environment, according to his access rights.

**Unified Data Management** process gives the ability to Data custodian to manage access policies and data across the data's lifecycle inside the data lake. As a part of data management, Data custodian also manages metadata. Metadata is defined centrally, using Metadata repository for storage of metadata, so that data can be found in the data lake and copied to sandbox environment when it is required by Business users for processing.

**Description of actors:**

**Business user** represents users across all levels and departments of an organization that uses Business Data Lake to gain insights from Big Data. Their goal is to use these insights towards achieving high-level drivers, objectives and goals of an organization.

**Data custodian** is responsible for data management and auditing in the Business Data Lake.

**Data subject** represents a person whose personal data is ingested and stored in the Business Data Lake.

**Description of business objects:**

**Data** represents raw data that is ingested into the data lake. It can be of various types, formats, level of structure and timelines.

**Distilled data** represents data with added structure. Structure to data is added through complex image processing, video analytics and graph and text analytical algorithms.

**Processed data** represents data that is processed according the specific needs that the Business user have, often for being an input to process of gaining insights.

**Actionable information** represents insights that can be used for making a business decision.

**Metadata repository** represents a register of metadata, that is to say data about data. Metadata is the information about data properties that is either embedded into the data file or it can be stored independently.

### 3.3.3 Develop Target Business Architecture Description

The problem that Target Business Architecture aims to solve is satisfying compliance requirements of GDPR, as postulated in chapter 3.3.11.

### *3.3.3.1 General view*

General view, shown as Figure 8, provides a high level abstraction of Business Data Lake's target business architecture. The inputs for diagrams' design are business scenarios, use cases and target requirements identified during Preliminary Phase and Architecture Vision Phase of TOGAF ADM.

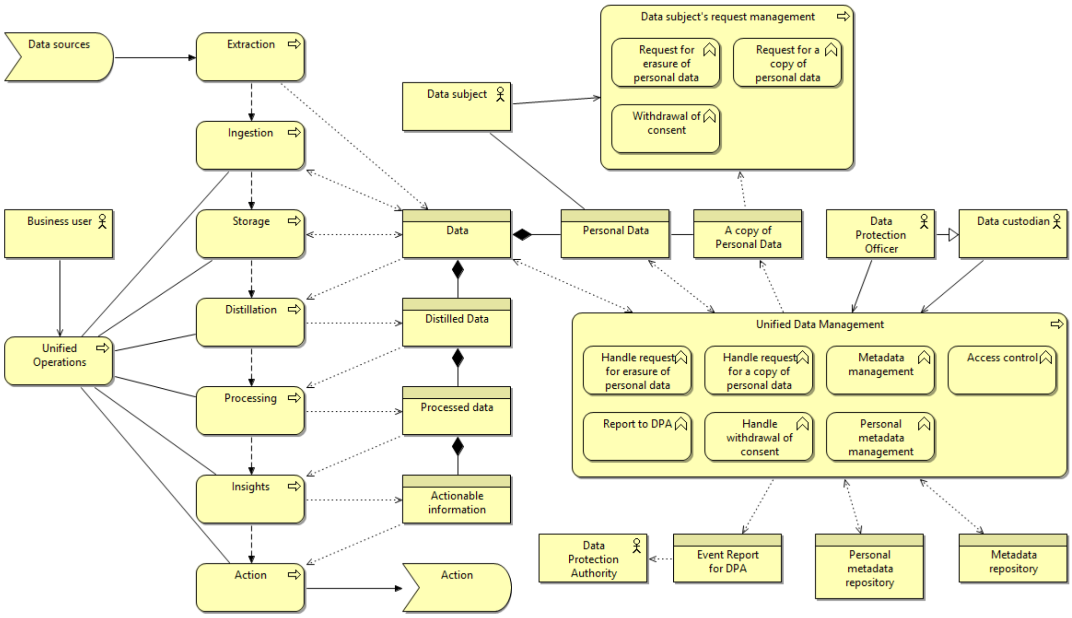


Figure 8, General view of target business architecture

## Process description:

**Data subject's request management** is a new process that enables a Data subject to create and manage the requests to an enterprise that cover Request for erasure of personal data, Request for a copy of personal data and Withdrawal of consent. A possible solution for this process would be a web service through which a Data subject could confirm his identity and create a request which is then delivered to the attention of Data Protection Officer. The officer would then be able to update the status of requests. Through the same service a copy of personal data could be safely delivered to the Data subject that made the request.

**Unified Data Management** process gives the ability to Data custodian to manage access policies and data across the data lifecycle inside the data lake. Metadata is defined centrally, using managed Metadata repository for storage of metadata, so that data can be found in the data lake and copied to sandbox environment when it is required by Business users for processing. New functionality in Target Business Architecture is maintaining a separate repository for personal data, reporting to DPA and handling requests for personal data erasure, copy of personal data and withdrawal of consent. This new functionality is managed by a new actor – Data Protection Officer. In order to enable Data Protection Officer to accomplish this task personal metadata management is implemented, supported with a personal metadata repository.

## Description of actors:

**Data Protection Officer** is a specialization of Data custodian. It is a Data custodian who is specialized for management and auditing of personal data and is responsible that these tasks are performed in compliance with GDPR.

**Data Protection Authority (DPA)** is a government agency tasked with control over enforcement of GDPR in enterprises.

## Description of business objects:

**Personal Data** represents data with an identified relation to a Data subject, which can be found in corresponding metadata.

**Copy of Personal data** represents a copy of personal data related to a specific Data subject that is prepared in response to a Request for a copy of personal data that was made by the same Data subject.

**Personal metadata repository** represents a register of metadata that identifies the data it describes as personal.

**Event report for DPA** represents the report that Data custodian/Data Protection Officer is obligated to create and send to Data Protection Agency after a security breach relating to personal data or another event relevant to personal data or data management that needs to be reported.

### 3.3.4 Perform Gap Analysis

Developed Business Architecture views showed that Baseline Business Architecture is not ready to support implementation of identified GDPR requirements, which motivated the design of Target Business Architecture. The purpose of Gap analysis matrix, shown in Table 2, is to identify and explain the gaps between Baseline and Target Business Architectures.

| | | Baseline Architecture | | |
|---|---|---|---|---|
| | | Unified operations | Unified Data Management | New processes |
| Target Architecture | Unified operations | Included | | |
| | Unified Data Management | | Potential match | Gap: To be enhanced |
| | Data subject’s request management | | | Gap: To be developed |
| | Eliminated processes | | | |

Table 2, Gap analysis matrix for Business Architecture

The gap matrix shows that Unified operations process, used by a Business user, has no gap as it remains unchanged between the baseline and target architectures.

Unified Data Management process is identified as a potential match. There is a gap between the baseline and target architectures, but it can be overbridged by design enhancement. Baseline functionality remains unchanged, while new functionality is added. The added functionality covers handling of Data subject’s requests that involve copying and erasing personal data. A necessary requirement for accomplishing these tasks is maintaining a separate repository for metadata that identifies data as personal. Another enhancement is added functionality for reporting to Data Protection Authority. The comparison between baseline and target state of Data custodian’s view of business architecture, together with Data subject’s view of target business architecture provides the design of needed enhancements.

Data subject’s request management is a completely new process that needs to be developed. Data subject’s view of target business architecture provides the design of this business process.

## 3.4 Information Systems Architectures

Phase C of TOGAF ADM is called Information Systems Architectures. It comprises of data and application architectures that are both developed at this phase (The Open Group, 2011).

### 3.4.1 Data Architecture

According to The Open Group (2011), the objective of Data Architecture is to enable Business Architecture and Architecture Vision.

#### *3.4.1.1 Select Reference Models, Viewpoints, and Tools for Data Architecture*

Data Architecture will be designed to support developed Business Architecture views. Data models described in Business Data Lake documentation will be used as a complementary information source in developing data architecture. The selection of views should follow the already developed views of Business Architecture.

Tools and techniques identified as appropriate for capture, modelling and analysis are business scenarios, use cases and baseline requirements that were produced during the Architecture Vision phase, as well as data architecture principles that were identified during the preliminary phase of TOGAF ADM.

Due to consideration towards readers the final version of this report will contain a detailed description of General view of Data Architecture, while only a gap analysis of other listed views will be included. Specialized views of Data custodian and Data subject are however available in great detail and can be provided upon request.

### *3.4.1.2 Develop Baseline Data Architecture Description*

#### 3.4.1.2.1 General view

General view of baseline data architecture enables general view of baseline business architecture, as shown on Figure 9.

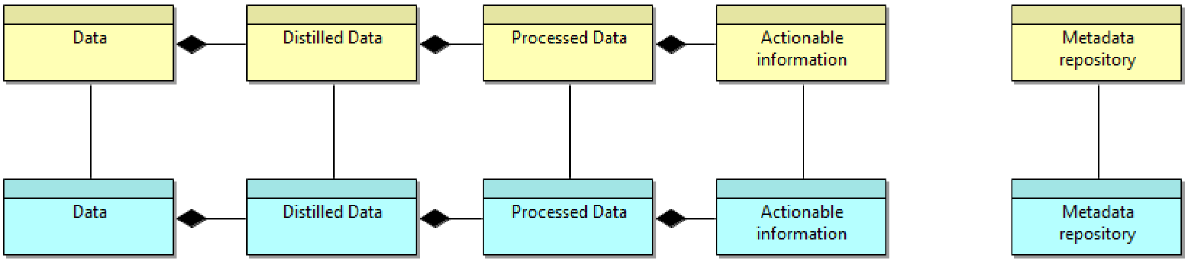


Figure 9, General view of data architecture baseline

**Data object description:**

**Data** object represents data that is ingested into the Business Data Lake and that comes in all types and levels of structure. It can be completely raw data or it can be highly structured. As of type, data can be picture, audio, video, text, csv, various document formats and many more. Level of metadata descriptiveness varies and capability for identification of personal data is inconsistent.

**Distilled data, Processed data and Actionable information** represent different stages that data goes through in processing sandboxes of the Business Data Lake. More structure is added at each stage, metadata is enhanced and sandboxed data is aggregated into new data that is released into storage as new Data when processing sandbox expires. Original data that was used as input into processing sandbox is never changed.

**Metadata repository** represents a key-value catalog where an address of data in the HDFS file system is paired with the corresponding metadata. The purpose of metadata repository is to enable efficient searching and management of data. A search for a keyword in metadata repository brings back the location of all data with metadata containing the keyword. A description of this process is shown on a Figure 10.

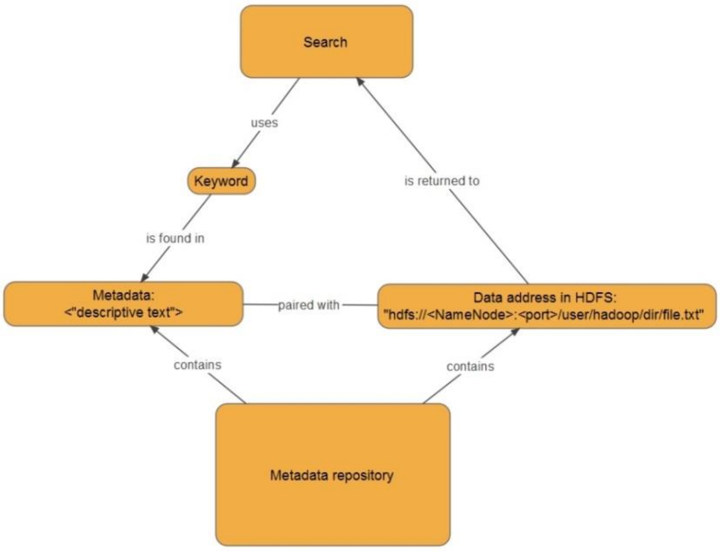


Figure 10, Operation principle of Metadata repository

### *3.4.1.3 Develop Target Data Architecture Description*

#### 3.4.1.3.1 General view

General view of target data architecture enables the general view of target business architecture, as shown on Figure 11.

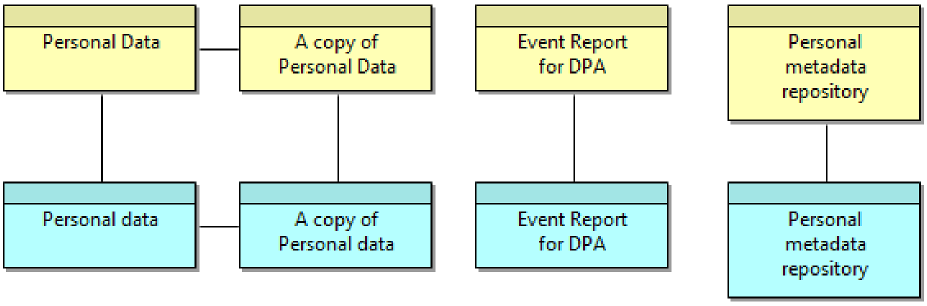


Figure 4, General view of target data architecture

<u>**Data object description:**</u>

**Personal data** object, as ordinary data object, represents data that is ingested into the Business Data Lake and that comes in all types and levels of structure. It can be completely raw data or it can be highly structured. As of type, data can be picture, audio, video, text, csv, various document formats and many more. The difference from ordinary data is that personal data is structured enough that its metadata contain the description of a person – a Data subject that the personal data relates to. The state when a personal data is ingested and stored without identification as such is theoretically a violation of GDPR and should be avoided as such.

**A copy of Personal data** represents an identical copy of Personal data.

**Event report for DPA** represents data in textual form.

**Personal metadata repository** represents a key-value catalog where an address of data in the HDFS file system is paired with the corresponding metadata that contains description of a person – a Data subject that the personal data relates to. The purpose of Personal metadata repository is specifically to enable efficient searching and management of Personal data. A search for a Data subject in metadata repository brings back the location of all Personal data belonging to that Data subject.

### *3.4.1.4 Perform Gap Analysis*

Data Architecture views showed that Baseline Data Architecture is not ready to enable Target Business Architecture, which motivates the design of Target Data Architecture.

| | | Baseline Architecture | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Data | Distilled data, Processed data and Actionable information | Metadata repository | Metadata | Personal data | Personal metadata | New data objects |
| Target Architecture | Data | Included | | | | | | |
| | Distilled data, Processed data and Actionable information | | Included | | | | | |
| | Metadata repository | | | Included | | | | |
| | Metadata | | | | Included | | | |
| | Personal data | | | | | Potential match | | Gap: to be enhanced |
| | Personal metadata | | | | | | Potential match | Gap: to be enhanced |
| | A copy of personal data | | | | | | | Gap: to be developed |
| | Event report for DPA | | | | | | | Gap: to be developed |
| | Personal metadata repository | | | | | | | Gap: to be developed |
| | Request and all its subtypes | | | | | | | Gap: to be developed |
| | Request summary report | | | | | | | Gap: to be developed |
| | Eliminated data objects | | | | | | | |

Table 3, Gap analysis matrix for Data Architecture

The gap matrix, shown in Table 3, identifies data objects that need to be enhanced or developed.

**Personal data** and **Personal metadata** exist per definition in baseline architecture state. However they are not identified and managed efficiently and consistently, which causes the situation where data and metadata that is personal exist without being structured and managed as such. In this regard, these data object have to be enhanced in order for target architecture state to be reached.

**Event report for DPA**, **Request and all its subtypes**, and **Request summary report** are data objects that do not exist in baseline data architecture and have to be developed from scratch. These objects have in common that they are textual in nature and in report form, so creating them would not present a significant challenge.

**A copy of personal data** is an identical copy of Personal data object. Once a Personal data object is enhanced and accessible, this gap would not prove difficult to close.

**Personal metadata repository** as such does not exist in baseline architecture state. Metadata repository exists, but the difference is that it is not dedicated to management of personal data and does not even contain consistent records of personal data in storage. Even if the records were consistent and current, each search for personal data would require a full search of the complete repository for yielding all personal data related to a single Data subject. Yet another problem is that not all instances (or versions) of personal data that physically exist on storage are registered in the repository. This prompts for the design of dedicated Personal data repository that would address all these issues.

### 3.4.2 Application Architecture

According to The Open Group (2011), the objective of Application Architecture is to, in collusion with Data Architecture, enable Business Architecture and Architecture Vision.

#### *3.4.2.1 Select Reference Models, Viewpoints, and Tools for Application Architecture*

Application Architecture will be designed to support developed Business Architecture and Data architecture views. Business Data Lake documentation will be used as a complementary information source in developing Application architecture. The selection of views should follow the already developed views of Business and Data Architectures. The abstraction level is not meant to detail the specific applications used for data processing and data management. Instead, application components will be described by their purpose. The reason for this design choice is the architecture principal that the solution should not be vendor-dependent. Another advantage of this approach is that the resulting enterprise architecture should be applicable not only on Business Data Lake, but on data lakes of other vendors as well.

Specialized views of Data custodian and Data subject are available in great detail and can be provided upon request.

#### *3.4.2.2 Develop Baseline Application Architecture Description*

##### 3.4.2.2.1 General view

General view of baseline application architecture, together with data objects, enables the general view of baseline business architecture, as shown on Figure 12.

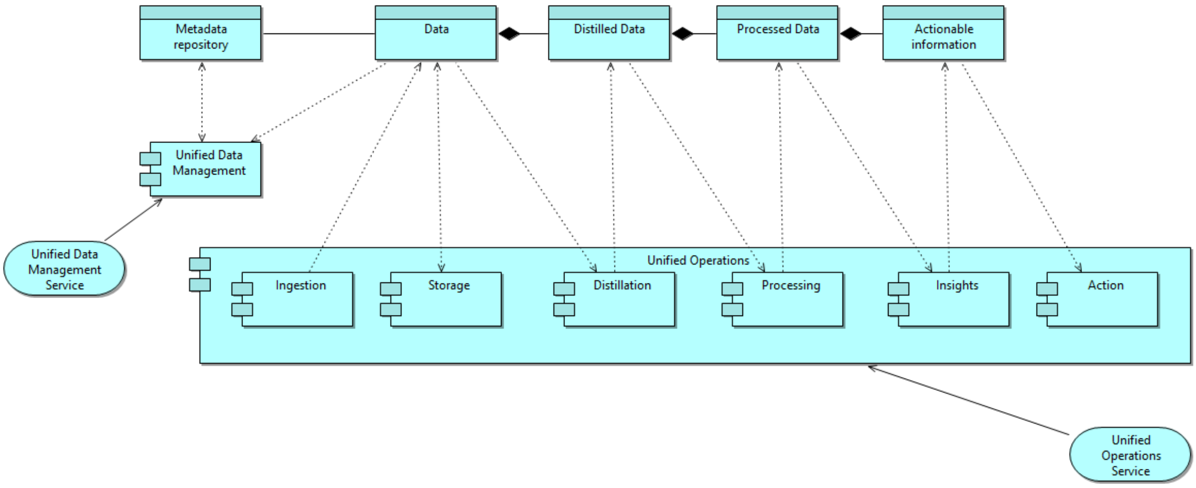


Figure 52, General view of application architecture baseline

**<u>Application description:</u>**

Ingestion, Storage, Distillation, Processing, Insights, Action, United Operations and Unified Data Management represent application components that together with Data objects enable Business processes of the same names that were described in the General view of Business Architecture Baseline.

Unified Operations Service and Unified Data Management Service are services that provide access to application components for their respective users.

### *3.4.2.3 Develop Target Application Architecture Description*

### 3.4.2.3.1 General view

General view of target application architecture, together with data objects, enables general view of target business architecture. Unified operations application has remained unchanged from baseline and will therefore not be described again, while Unified Data Management developed new functionality, as shown on Figure 13.

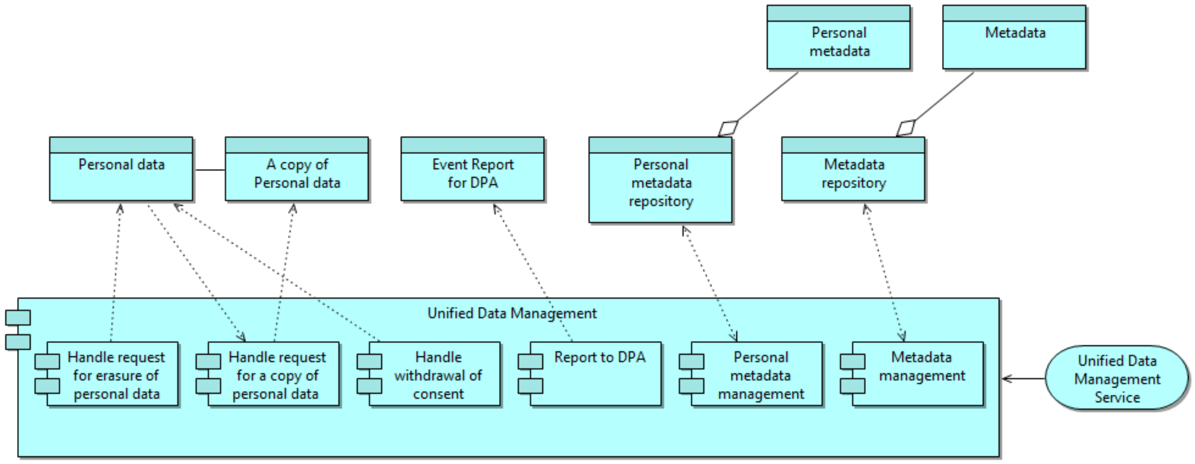

Personal metadata
Metadata
Personal data
A copy of Personal data
Event Report for DPA
Personal metadata repository
Metadata repository
Unified Data Management
Handle request for erasure of personal data
Handle request for a copy of personal data
Handle withdrawal of consent
Report to DPA
Personal metadata management
Metadata management
Unified Data Management Service

Figure 13, General view of target application architecture

**Application description:**

Unified Data Management application has changed significantly. New sub-components include:

- **Report to Data Protection Authority** application component, which together with data objects enables Report to Data Protection Authority business process from Target Business Architecture general view.
- **Handle request for a copy of personal data** application component, which together with data objects enables Handle request for a copy of personal data business process from Target Business Architecture general view.
- **Handle request for erasure** of personal data application component, which together with data objects enables Handle request for erasure business process from Target Business Architecture general view.
- **Handle withdrawal of consent** application component, which together with data objects enables Handle withdrawal of consent business process from Target Business Architecture general view.
- **Personal metadata management** application component, which together with data objects enables Personal metadata management business process from Target Business Architecture general view.
- **Metadata management** application component, which together with data objects, enables metadata management business process from Target Business Architecture general view.

### *3.4.2.4 Perform Gap Analysis*

As Application Architecture views show, Baseline Application Architecture is almost completely unprepared to enable Target Business Architecture, which motivates the design of Target Application Architecture.

| | | Baseline Architecture | | | |
|---|---|---|---|---|---|
| | | Unified Operations Application | Pivotal Data Dispatch | Unified Data Management | New applications |
| Target Architecture | Unified Operations Application | Included | | | |
| | Pivotal Data Dispatch | | Included | | |
| | Unified Data Management Application | | | Potential match | Gap: To be enhanced |
| | Personal metadata management | | | | Gap: to be developed |
| | Report to DPA | | | | Gap: to be developed |
| | Handle withdrawal of consent | | | | Gap: to be developed |
| | Handle request for a copy of personal data | | | | Gap: to be developed |
| | Handle request for erasure of personal data | | | | Gap: to be developed |
| | Data subject’s request management | | | | Gap: to be developed |
| | Request for a copy of | | | | Gap: to be developed |

| | | | | | |
|---|---|---|---|---|---|
| | personal data | | | | |
| | Request for erasure of personal data | | | | Gap: to be developed |
| | Withdrawal of consent | | | | Gap: to be developed |
| | My files | | | | Gap: to be developed |
| | Create account | | | | Gap: to be developed |
| | Login | | | | Gap: to be developed |
| | Eliminated applications | | | | |

Table 4, Gap analysis matrix for Application Architecture

The gap matrix, shown in Table 4, identifies application components that need to be enhanced or developed in order to enable Business architecture.

**Unified Data Management application** must be enhanced significantly with a line of application sub-components that are completely missing in baseline application architecture:

- Personal metadata management
- Report to DPA
- Handle withdrawal of consent
- Handle request for a copy of personal data
- Handle request for erasure of personal data

**Data subject's request management** application is completely missing in baseline application architecture, together with its sub-components:

- Request for a copy of personal data
- Request for erasure of personal data
- Withdrawal of consent
- My files
- Create account
- Login

The gap analysis of the application architecture showed that there are many serious gaps, that is to say many applications that need to be developed in order to make Business Data Lake compliant with GDPR.

## 3.5 Technology Architecture

According to The Open Group (2011), the objective of Technology Architecture is to enable the logical and physical application and data components, as well as Architecture Vision.

### 3.5.1 Select Reference Models, Viewpoints and Tools for Technology Architecture

Technology Architecture will be designed to support developed Information Architectures. Due to the unsuccessful interview with a technology expert, Business Data Lake documentation will be used as a primary information source in developing Application architecture. As a complementary source, technology articles and interviews with experts in other areas will be used respectively. The selection of views should follow the already developed views of Business and Data Architectures. However the specific technology used for enablement of Information architecture may prove to be unchangeable as no viable replacement exists, requiring only one view. Additional inconvenience emerged with the failure of the interview with technology expert that was meant to provide a more detailed perspective into Technology Architecture of Business Data Lake.

Due to consideration towards readers the final version of this report will contain a detailed description of General view of Technology Architecture, while only a gap analysis of other listed views will be included. Specialized views are however available and can be provided upon request.

### 3.5.2 Develop Baseline Technology Architecture Description

General view of baseline technology architecture enables the general view of baseline information architectures, as shown on Figure 14. Baseline technology layer is entirely based on Hadoop open source software framework for scalable, distributed, data-intensive computing. For the reason that the same technology configuration enables all baseline views, only general baseline view is designed.

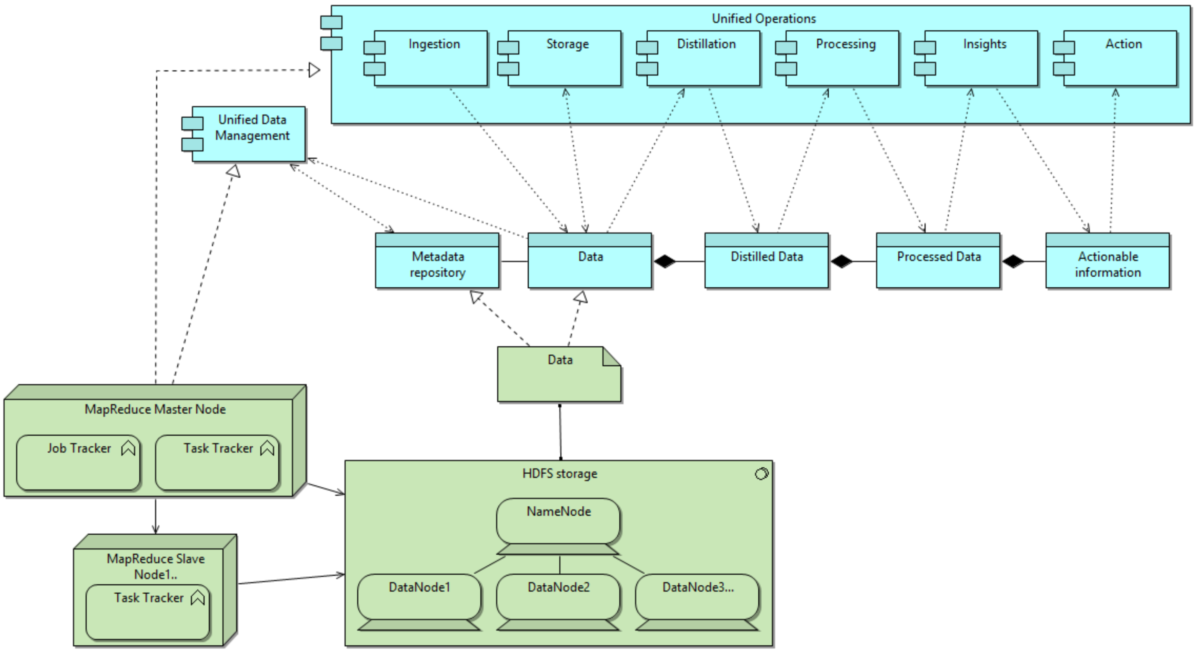


Figure 14, General view of baseline technology architecture

**Technology description:**

**Data** artifact is a physical manifestation of data. It realizes data and metadata data objects as files in physical storage.

**HDFS storage** stands for Hadoop Distributed File System. It is a scalable, distributed storage that enables high-throughput access to application data. HDFS cluster consists of:

- **NameNode** that contains file system metadata
- **DataNode** that contains the actual data

HDFS is designed for reliability and fault tolerance. This is achieved by replicating data across multiple hosts – DataNodes. Files in HDFS are divided into blocks that are stored across DataNodes in a large cluster.

**MapReduce** is a software framework that facilitates distributed processing of Big Data in parallel on large computer clusters. **MasterNode** accepts a processing job from application layer and divides the job into components that are then scheduled to be processed by **SlaveNodes**.

**JobTracker** function schedules the job's component tasks on SlaveNodes, monitors their progress and restarts failed tasks.

**TaskTracker** function executes the task that node was scheduled by MasterNode through JobTracker.

## 3.5.3 Develop Target Technology Architecture Description

### *3.5.3.1 General view*

General view of target technology architecture enables the general view of target information architectures, as shown on Figure 15. The most significant change from baseline architecture is External communication node, which enables information exchange between Data subject and Data Protection Officer.

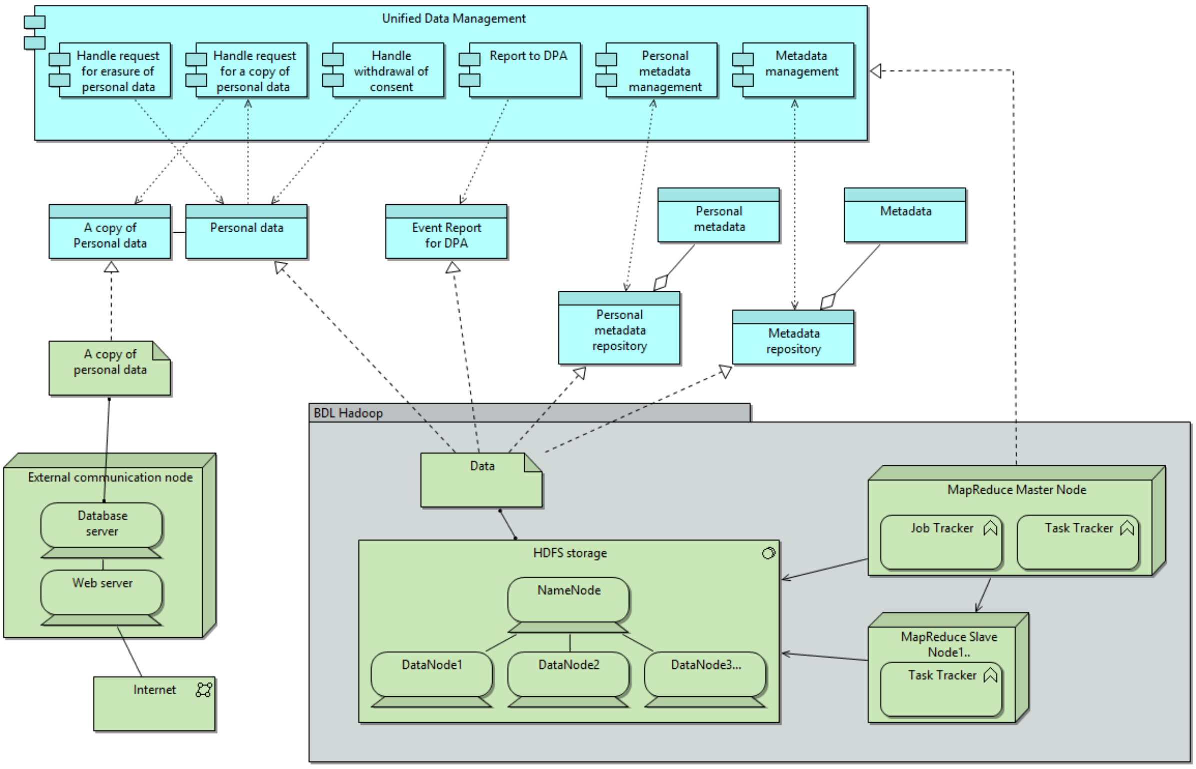


Figure 6, General view of target technology architecture

**<u>Technology description:</u>**

**Data** artifact is a physical manifestation of data. It realizes data and metadata data objects as files in physical storage. Data object of target architecture; Personal data, Personal metadata repository and Event report for DPA, are also realized as Data on HDFS storage.

**A copy of personal data** artifact realizes a copy of Personal data object that is realized as Data artifact in HDFS storage. A copy is realized on external system that is dedicated to information exchange with data subjects. The reason for this configuration is to provide a stronger security by limiting external access to Business Data Lake system.

**External communication node** represents a separate system that is used for information exchange with data subjects. The system consists of a **Database server** that stores the information to be exchanged and of a **Web server** that provides external access for data subjects. Database server, as well as Web server, can be realized through a variety of open source solutions, avoiding vendor dependency.

The rest of technology layer has remained unchanged from baseline architecture.

### 3.5.4 Perform Gap Analysis

Developed Technology Architecture views show that there is a difference between the baseline and target technology architectures. Identified gaps are listed in the gap matrix table:

| | | Baseline Architecture | | | | |
|---|---|---|---|---|---|---|
| | | HDFS storage | MapReduce Master Node | MapReduce Slave Node | Data | New technology infrastructure |
| Target Architecture | HDFS storage | Included | | | | |
| | MapReduce Master Node | | Included | | | |
| | MapReduce Slave Node | | | Included | | |
| | Data | | | | Included | |
| | External communication node | | | | | Gap: to be developed |
| | Request | | | | | Gap: to be developed |
| | A copy of personal data | | | | | Gap: to be developed |
| | Request summary report | | | | | Gap: to be developed |
| | Eliminated technology infrastructure | | | | | |

Table 5, Gap analysis matrix for Technology Architecture

The gap matrix, shown in Table 5, identifies technology infrastructure that needs to be enhanced or developed in order to enable Information Systems (Application and data) architectures.

**HDFS storage and MapReduce Master/Slave Nodes** remain unchanged between baseline and target technology architecture.

**Data** artifact remains unchanged as well. Although there are new artifacts in target architecture that are its subtypes, Data already in baseline has the property to support all data types.

**External communication node** represents technology infrastructure that facilitates information exchange between Data subject and Data Protection Officer. Such functionality could also be achieved by enhancing Hadoop infrastructure, but for security reasons it is judged as best to opt for a separate piece of infrastructure. Facilitating such functionality is a rather common task for technology infrastructure, so this gap cannot be considered as significant or hard to overbridge.

**Request** is a new artifact that enables the Request data object. It is presumed to be a text file type.

**A copy of personal data** is a new artifact that enables the data object of the same name. A challenge for this gap to be overbridged is that a variety of file types that are stored in Business Data Lake must be transformed to a format that is easily interchanged between different systems. Open source formats should have priority when an appropriate format is chosen.

**Request summary report** is a new artifact that enables the data object of the same name. It is presumed to be a text file type.

## 3.6 Requirements Management

During the design of Baseline and Target architectures, and the subsequent gap identification and analysis, additional requirements were captured. These requirements are validated against previously captured requirements and architecture principles, and are available for review in Appendix J (see Tables J1 – J14). A certain amount of redundancy is possible as validation was focused primarily on preventing possible collisions between requirements.

# 4 Compliance Solution Proposal and Evaluation

During the course of the project so far, separate views were designed inside each layer of both the baseline and the target architecture. Separation of views enabled presentation and analysis of enterprise architecture on different levels of abstraction. The reason for designing separate views is that using one single view for visualization of the whole architecture, or even a whole layer, would not be optimal due to complexity of the model. Furthermore, by using different levels of abstraction and scope, separate views provide context to visualized architecture, which makes both understanding and analysis of architecture easier.

The result of the performed design work is embodied in the presented architecture design and gap analysis that is summarized in Table 6 which will be used to answer the research questions and lead to a solution proposal.

| | Gap source | Action needed | Value proposal | Impact |
|---|---|---|---|---|
| Business Architecture | Unified Data Management | To be enhanced | Added functionality towards GDPR compliance includes handling of Data subject's requests that involve copying and erasing personal data, as well as withdrawal of consent. Managing a separate repository for personal metadata and reporting to Data Protection Authority are also added. | High |
| | Data subject's request management | To be developed | Data subject can create and manage requests for exercising the rights given to him by GDPR. The functionality also covers managing information that is replied by the Data Protection Officer. | Moderate |
| Information Systems Architecture | Personal data | To be enhanced | All personal data that is ingested and stored should be identified and managed as such. | High |
| | Personal metadata | To be enhanced | All personal metadata must be identified and collected in order for personal data to be recognized. | High |
| | A copy of personal data | To be developed | Represents a copy of personal data object that is needed to satisfy the Right to a copy of personal data GDPR requirement. | Low |
| | Event report for DPA | To be developed | Represents a data object needed to satisfy Report to Data Protection Authority GDPR requirement. | Low |
| | Personal metadata repository | To be developed | Represents a means to address and locate every piece of personal data connected to a specific Data subject. | High |
| | Request and all its subtypes | To be developed | Represents the request data object that Data subject creates and sends to DPO in the hope of fulfilling one of the rights he has been given by the GDPR. | Low |
| | Request summary report | To be developed | Represents a data object that is created as a report on handling specific request. | Low |
| | Unified Data Management | To be enhanced | Application that is used by Data custodian and Data Protection Officer to | High |

|  |  |  |  |  |
|---|---|---|---|---|
|  | Application |  | manage data and access policies. The enhancement brings separate management of personal data, management of Data subjects' requests, and the capability to report to Data Protection Authority. |  |
|  | Personal metadata management | To be developed | Application that enables Data Protection Officer to separately manage metadata that describes ownership of personal data. Should facilitate more consistent and reliable management of personal data. | High |
|  | Report to DPA | To be developed | Application component that that enables Data Protection Officer to create and send a report to Data Protection Authority. | Low |
|  | Handle withdrawal of consent | To be developed | Application component that enables Data Protection Officer to erase personal data per request of Data subject and regulate access to subject's personal data. | Moderate |
|  | Handle request for a copy of personal data | To be developed | Application component that enables Data Protection Officer to create a copy of personal data per request of Data subject. | Moderate |
|  | Handle request for erasure of personal data | To be developed | Application component that enables Data Protection Officer to erase personal data per request of Data subject. | Moderate |
|  | Data subject's request management | To be developed | Represents application that will be used by Data subjects for creating and managing the requests that are sent to DPO and their replies, including copies of personal data. | Moderate |
|  | Request for a copy of personal data | To be developed | Application component that enables Data subject to create and send Request for a copy of personal data. | Low |
|  | Request for erasure of personal data | To be developed | Application component that enables Data subject to create and send Request erasure of personal data. | Low |
|  | Withdrawal of consent | To be developed | Application component that enables Data subject to create and send Request for a copy of personal data | Low |
|  | My files | To be developed | Application component that enables Data subject to receive, store and review communication and files that are sent to that particular Data subject by a Data Protection Officer. | Low |
|  | Create account | To be developed | Application that enables Data subject to create a new user account. | Low |
|  | Login | To be developed | Application that enables Data subject to login to existing user account. | Low |
| Technology Architecture | External communication node | To be developed | A separate system used for communication and information exchange with external actors – Data subjects. | Moderate |
|  | Request | To be | Physical manifestation of Request data | Low |

|  |  |  |  |  |
|---|---|---|---|---|
|  |  | developed | object and its subtypes. A text file. |  |
|  | A copy of personal data | To be developed | Physical manifestation of A copy of personal data object, in an interchangeable file format. | Low |
|  | Request summary report | To be developed | Physical manifestation of Request summary report data object. A text file. | Low |

Table 6, Gap analysis matrix for Enterprise Architecture

Closing all of the gaps that are listed in Table 6 is necessary in order to achieve the compliance with GDPR. The level of impact for each gap shows the significance of closing the gap towards reaching compliance.

Being identified as the gap with highest impact on compliance with GDPR, solution design for Personal metadata repository will be proposed in the next chapter.

## 4.1 Solution Design for Personal metadata repository

Following gap analysis on different architectures, the gap whose closing would create the most impact on compliance with GDPR is identified as Personal metadata repository. In order for a solution to be proposed, one must first describe the desired functionality of the target architecture.

Personal metadata repository represents the means to address and locate every piece of personal data connected to a specific Data subject. In baseline architecture, there is no metadata repository that is dedicated to personal data. Instead, there is a metadata repository that contains all metadata available on all data. In this repository, metadata and the location of data in storage are connected through a key-value pair that consists of metadata and data’s address in HDFS (see figure 16). Key-value pairs are organized into tables that are profiled to service the searches that need to be performed. This solution enables the search for personal data, under the condition that such a key-value pair table is created and that personal data is systematically and consistently identified (which is not the case today). Apart from the problem with identification of personal data, which can only be solved in coordination with data sources, there is a large room for improvement in managing the personal data.

Using a table as a metadata registry presents certain issues in the Big Data environment. Obviously, the table will be huge. Other issues, which are related to the way Hadoop stores files, are versioning and redundancy. In order to achieve GDPR compliance in managing all the personal data, all instances of this data must be managed as well. Even old, unused versions of files that contain personal data have to be managed. This is not the case with the current configuration of metadata repository and even if it were, including all existing instances of data into a table would greatly multiply the number of items in that table. This would make the table overly complicated and difficult to manage, while searching for personal data related to a specific data subject in such a table would not be optimal because the full search would have to be made, checking all the items that the table contains.

The solution to the problems described above is using a graph database for storing key-value pairs that identify the location of personal data. Personal metadata, that represent a Data subject, and the address that identify the data's location should be stored as nodes in a graph, connected with edges. The same should be done with all versions and instances of data, with metadata labeled accordingly.

The graph database used for this design is Neo4j, due to availability of the database and expert support that could be consulted in person inside the time-frame for this thesis. It is however important to note that the solution is not limited to Neo4j in particular, as the model can just as well be used with other graph databases, such as OrientDB.

The first step in designing a graph database that would enable metadata repository is to establish entities, their relations, and the properties they might have, as shown in Figure 16.

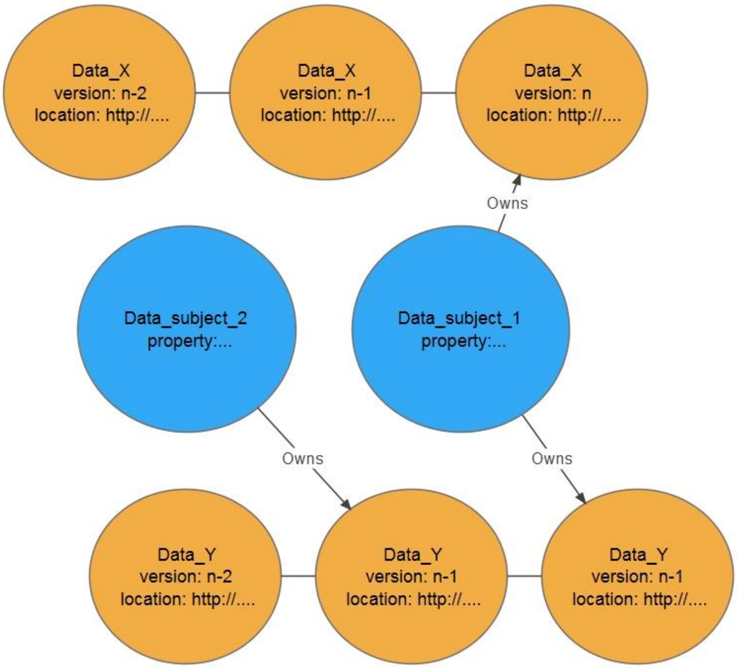


Figure 16, Graph model of Personal metadata repository

Figure 16 shows a graph model representing Data subjects and Metadata entities as nodes. Data subject node's identifier is Data subject's name. The node supports properties that can be used for various purposes, such as to describe the data subject, input contact information, input information about consent, or any other use that Data custodian/Data Protection Officer might have. Data subject nodes are linked by edges to metadata that describes the data relating to respective Data subject.

Metadata entities are identified by the same name as the data they describe. Other properties include version number and the location of the instance of data that the metadata describes. Metadata is linked by edges to metadata that represents other instances or versions of the same data.

As Figure 16 shows, Data_subject_1 is related to Data_X and Data_Y, including all previous versions of that data. Data_subject_2 is related to a previous version of Data_Y, as relations may appear and disappear while data is processed. It is important to note that edges can have properties as well and that they can be used to describe the nature of the relationship, such as if the relationship transfers to other connected nodes or if it is limited to that particular node.

The main advantage of this solution is that query time is reduced drastically. Every node in a graph database physically contains a list of the nodes it is connected to. This eliminates the need for expensive JOIN operations running search/match computation that are used in traditional relational databases. When, for example, Data Protection Officer needs to locate all the data related to Data_subject_1, the query accesses the Data_subject_1 node, reads the list of nodes that are connected to it and directs access to connected nodes. According to Robinson et al. (2013), the gain in performance is that of several orders of magnitude.

Another advantage is in the obvious simplicity of the design and ease of maintenance in comparison to traditional RDBMS solutions.

There is however one requirement that must be satisfied for this design to be successful in regards to the compliance issues in the management of personal data; personal metadata must be captured consistently and efficiently as data is ingested and stored into the data lake. As this is the requirement that must be met in cooperation with data sources, it is outside the scope of this work. However, a recommendation can be made that in such a case that the flow of metadata from data sources is regulated in a satisfactory way, an application for populating the personal metadata repository should be designed. An alternative, and a valid example for the functionality that is required, would be acquiring an external solution, such as Revelytix Loom. This application, according to Davis, B. (2014), supports an extensible metadata registry, active scan for capturing metadata, keeping track of data transformations and lineage, and can be integrated with existing platforms and tools – all in Hadoop.

As a final step of the design performed during this thesis, graph implementation of Personal metadata repository was prototyped in Neo4j graph database. Screenshots and construction code in Cypher query language are available for review in Appendix K, while examples and results demonstration of three typical queries are available in Appendix L.

The prototype consists of two different types of nodes that are distinguished through labels:

1. **DataSubject**, of which a 100 nodes with properties ID and name. ID is a unique number from 1-100, while name was populated through repetition of 10 different names in combination with ID number in order to make the name distinguishable.
2. **Address**, of which 50 nodes have a secondary label V1 that denotes the version of the file that is stored on that address, 50 nodes have a secondary label V2, and another 50 nodes have a secondary label V3. Properties of Address nodes are unique ID, unique name that points to the address the file is stored at, and version that gives information about the version of the file that is stored at pointed address.

One hundred random relationships (edges), which are labeled as :OWN, are created between DataSubject and Address nodes, directed from DataSubject to Address nodes, in order to simulate the link between Data subjects and their data. An option of a single data instance containing the personal data belonging to multiple Data subjects is permitted.

The prototype effectively demonstrates the functionality of a graph based Personal metadata repository, by servicing queries that enable locating all instances and versions of data belonging to a specific data subject, as well as all or specific Data subjects that are related to particular data.

Differentiation of query results is possible through labels, node properties, relationships (edges in graph terminology) or relationship types. The prototype even enables running of more complicated queries, such as determining if a located instance of data is shared between several data subjects, or other analytical purposes. Again, examples and results of abovementioned queries are available in Appendix L.

As the prototype is intended only to demonstrate the design and functionality of Personal metadata repository, only properties that are crucial to this purpose are added to the model. Nevertheless, it is reasonable to assume that more descriptive properties of DataSubject and Address nodes, as well as that of their relationships, can be added and used to enable even more complex queries and analysis. It is even apparent that these additions have the potential to broaden the application area of the repository, from being limited to management of personal data to analytics in general.

## 4.2 Evaluation

The project as a whole, including both the method workflow and the achieved results, as described in this report, were presented to and evaluated by the project's sponsor who provided the following comments:

Chosen methods, namely TOGAF ADM, were evaluated as appropriate for solving the given problem. Business scenarios of both the baseline and the target architecture vision were correctly described on a high level of abstraction. Management of personal data and particularly enabling the satisfaction of GDPR requirements which relate to Data subjects' rights over their own personal data were correctly recognized as crucial in reaching compliance with GDPR.

Designed Enterprise Architecture is evaluated as satisfactory. Chosen views are relevant to description of architecture areas that are targeted by new requirements. Description of views provides the needed context. Both the baseline and the target architecture are designed and presented in a way that is easy to understand.

Gap analysis presents a clear picture of compliance issues that baseline architecture has with GDPR and of the scope of changes that need to be done in order for compliance to be reached. Parts of enterprise architecture that remain unchanged, that could be improved, as well as parts that need to be developed altogether, are described very visually and can easily be traced to Enterprise Architecture views.

The designed prototype for personal metadata repository is an unexpected bonus value of this project. The functionality of the prototype is demonstrated, both with the attached code and screenshots showing the results of typical queries. The solution is evaluated as simple, effective and with the potential to be implemented in the Business Data Lake, as well as other solutions for data processing and analysis. Commercialization of the solution as a standalone product can also be considered as an avenue of future research.

The overall evaluation of the project is that it fulfilled the expectations of the project's sponsor in terms of establishing the lack of Business Data Lake's compliance with GDPR and, with restrictions, determining changes that would be necessary to design in order to achieve the compliance. The project even exceeded the expectations in that it yielded a fully functional prototype of a personal metadata repository, which is correctly identified as a key gap that enables bridging of other gaps that were identified during the project's course.

As a remark, project's sponsor states that architecture views could be broken down into more detailed views of a lower abstraction level that would perhaps provide a richer base for gap analysis and lead to a more detailed description of changes. Particularly, Technology architecture could be designed in greater detail that would include an important compliance issue known as immutability of Hadoop. Furthermore, project's sponsor expressed the desire for a more compact version of the report, with more focus on design and results, rather than on method.

Project sponsor also notes that the project does not cover all the requirements of GDPR and concurs that some requirements may change over the course of GDPR's finalization. More importantly, the designed solution relies on solving at least three very serious issues regarding the compliance with GDPR. The first issues is the inability to permanently delete specific data instance in Hadoop, while the other issue is regulation of the ingestion process in a way that incoming data is reliably screened for personal data. The third issue is the need for permanent screening solution for personal data inside the data lake, which is necessary for the reason of the dynamic nature of personal data – ordinary data can through processing and transformation during analysis also become personal. For this reason, the designed architecture, even in high abstraction level, does not deliver the complete design solution for compliance of Business Data Lake with GDPR.

Despite the given remarks, project's sponsor considers the project successful and accomplished results a solid base for continuance of work towards achieving compliance of Business Data Lake with GDPR.

# 5 Discussion

During the course of this work, several interesting discussion topics emerged. These topics are related to the way the work was done, design decisions, but also to diverging opinions about different aspects of the technology that was, or could be used in the solution design.

## 5.1 The difference between working on an EA model and on a deployed EA

It is important to note that the design work performed during the course of this thesis is done on non-deployed model of Business Data Lake. Deployed Business Data Lake may differ from a model that was presented here, for the reason of being adapted to suit a specific application area.

It is also important to understand that specific applications, such as health data processing, historical, statistical, and scientific research that represents public interest may present an entirely different set of compliance requirements for management of personal data. These requirements may be stricter, but they can also be more relaxed, in order to protect more sensitive data or to enable processing of data that represents public interest.

Such specializations are separate cases that require adaption of a BDL architecture model at deployment. The objective of this work was not to provide a universal out-of-the-box solution for all possible applications for data lakes. Such a design would be unnecessarily complicated and unfeasible for a number of reasons, which are embodied in the architecture principle BP02 (see Appendix C) that states that changes to applications and technology are made only in response to legitimate business needs. Some specialization cases may even be in a direct collision with one another. For these reasons, the design work is performed on a non-deployed architecture model of BDL, allowing for versatile design that can serve as a base for eventual future specializations.

## 5.2 Synchronization of Methods and Applying Iteration

Using a total of three methods simultaneously throughout the course of this project may appear overly complicated at first glance. However, after completing the project it is difficult to imagine another setup that would have been more optimal.

Applying iteration to methods, during the course of this project, can be compared to designing clockwork. Each of the three methods operates on a separate iterative layer, with overlapping between iterations:

1. Design research, as scientific method, represents the top iterative layer.
2. TOGAF ADM, as design method, represents the middle iterative layer.
3. Open Kanban, which is used as workload management method, represent the bottom iterative layer.

Design Research scientific method, was largely satisfied through the use of TOGAF ADM as design method. Design Research cycles were iterated, as planned, both to TOGAF ADM as a whole, as well as on phases and phase groups, in order to gain new knowledge through design. Under the constraints of this project, not all phases of Design Research could be directly mapped to TOGAF ADM or its phases. Plan, Observe, and Design phases of Design Research method could, over the greater part of the project, be iterated to the first five phases of TOGAF ADM, both individually or as a whole. Including Prototype and Test phases to iterations became possible after the design of enterprise architecture was pretty much finished. Starting with the solution design of Personal metadata repository added Prototype phase to iterations of Design Research method. Test phase was added to iterations somewhat unconventionally, through Project sponsor's evaluation of the design.

Open Kanban's actual purpose is to be used as a method for effective cooperation inside a project group. Under the constraints of this project, Open Kanban was used very effectively as workload management method. Performing the steps of TOGAF ADM phases as batch sizes in Kanban made it possible to follow the status of both the individual steps and of the whole project.

As Kanban board, an ordinary Plexiglas whiteboard was used. The reason for choosing this solution over the freely available software solutions is purely subjective. Although the software enables many useful functions, such as calculating the time to project's completion, analysis of the progress through diagrams, or readily sharing of the Kanban board, its use would require another open window on an already overpopulated computer screens. The course of this project required simultaneous use of several software applications, and many open windows on two computer screens. Any chance of avoiding any unnecessary distraction on screen was readily welcomed.

Mapping steps of TOGAF ADM phases as Kanban batch sizes proved to be an optimal choice for granulation of workload increments for this project. No step was found to be a poor match for being managed as batch size, which may create a tendency towards using the same setup for the continuance of the project. Such an approach must be carefully considered because this project was performed on a relatively high level of abstraction. A lower level of abstraction usually means more details, which in turn can cause the rise of the amount of workload per step. In such a case, managing steps of TOGAF ADM as Kanban batch sizes could not be considered optimal, prompting for further granulation of workload.

## 5.3 Description of links between different states of data in the Business Data Lake

While being processed inside the Business Data Lake, the data goes through different states on its way to actionable information as shown on general views of both the baseline and target architectures. Data that is stored in HDFS is transformed first into distilled data, then to processed data, in order to finally be transformed into actionable information. These processes are performed in a sandbox environment, which, depending on the data's type and intended usage, may be in physical or virtual memory. Data transformation may not go all the way and can finish at distillation or processing, but in all cases, the newly created data is returned into HDFS when it's no longer actively needed.

A problem emerged while designing views of business architecture, regarding how to describe the links between different states of data. Distilled, processed data and actionable information derives from data and is later returned to HDFS to become data themselves. The problem that occurred is how to connect these dynamic states of data on a static view.

While data is processed, the derived states of data should be represented as specialized subtypes of data, requiring to be connected by a specialization relation. On the other hand, when the derived data is returned to HDFS, it becomes a part of data, requiring a composition relation.

This problem is much more complex than it seems, because the choice of a relation type ultimately determines if the derived data that comes out of sandboxes falls under the scope of personal data management before or after it comes out. In the first case, derived data should be represented as a specialization of data because it requires separate process for personal data management, while in the latter case personal data management occurs in the same way for any other data that is ingested and stored in HDFS. Therefore, a seemingly unimportant choice between two equally applicable relation types in reality determines how an entirely different process is meant to operate.

A design choice is made in this case that all personal data management should occur when the data is ingested and stored in HDFS and that returning of transformed data that comes out of sandboxes should be considered as ingestion, which is represented on architecture views by using composition relation between data and its derivatives.

## 5.4 Immutability of HDFS

Immutability is a property of HDFS that is the other side of the medal that is reliability and fault-resistance.  This property prevents the files in HDFS to be changed or deleted. Instead, a new instance of the file is created (free from the data that was deleted) and the latest instance automatically becomes an active one. Previous instances of files are not deleted, only flagged as not active. This presents a problem for achieving compliance in relation to deletion of data, because the data remains undeleted in the previous instances of HDFS files.

The passage above is a much simplified explanation of Hadoop immutability issue. The issue was placed out of the scope of this project for two reasons.

The first reason is that this is a serious issue that targets all Hadoop based architectures, not just data lakes. It is also connected to the very nature of how Hadoop operates, and changing it may prove to be a very complicated matter. By its broad area of impact, as well as the possibility of taking considerable resources to address, this issue is out of the scope of the project as this project is limited both in time and resources.

The second reason is that the information, that was necessary for detailed understanding of Hadoop inner workings inside the Business Data Lake, failed to be acquired. An interview that was planned as source of information for this particular area did not succeed because of the unavailability of the intended correspondent. Time-limit for completion of this project proved to be too tight for finding another correspondent of equivalent expertize, which led to discussions of this issue with other two correspondents. The discussions revealed some aspects of the problem, but that was not enough to model it on designed architecture views.

The fact that Hadoop immutability problem is not present on the list of gaps between baseline and target technology architectures could be seen as misleading. The reality is that this is a very serious issue that is critical to compliance with GDPR because it determines if the deletion of data can be performed in accordance with regulation.

## 5.5 Anonymization

As described under the Theory section of this thesis, there are conflicting opinions about usability of anonymization in protection of personal data. Some opinions, such as that of Sedayao et al. (2014), advocate that anonymization is a viable solution for processing of personal data, because it makes the data not personal and therefore not subject to laws that regulate the processing of personal data. Others, such as Ohm (2010) as well as Narayanan and Shmatikov (2008), believe that anonymization cannot guarantee that re-identification of data as personal will not occur.

A design choice was made not to rely on eventual benefits of anonymization in achieving compliance with General Data Protection Regulation. The reason for this decision is that relying on anonymization comes with a constant risk of finding oneself in a costly breach of compliance with GDPR, if re-identification of personal data occurs. Assessing the risk of re-identification is a very complex and uncertain process. Furthermore, one of the key points of GDPR states that security should be a part of the design from the start, not an afterthought. For all those reasons, a choice was made to design a target solution that would be compliant with GDPR without being susceptible to re-identification risks that follow anonymization of data.

## 5.6 Gartner's critique of Data Lakes

White and Heudecker wrote a report in 2014 on behalf of Gartner, criticizing the data lake concept as flawed because of the gaps in semantics, governance and security, while advocating the concept of data warehouses as the way of the future (White and Heudecker, 2014). Harper (2014) largely agrees with this critique, naming metadata and semantics as some of the areas in which data lakes should be improved.

While working on baseline architecture views, problems that were mentioned in Gartner's report were encountered. Through the development of target architecture and subsequent gap analysis, it became clear that many of the identified gaps coincide with the gaps described by White and Heudecker.

However, conclusions of Gartner's report differ from the first-hand experience of data lakes that was acquired through performing this project. It is true that the results show that baseline architecture is plagued by the flaws described by White and Heudecker. It can even be said that these flaws are the primary cause of the compliance problems that the baseline architecture has with GDPR. The point where the conclusion differs from Gartner's report is that the design of target architecture shows that these flaws are manageable and that there is a way to improve the management of metadata and semantics in a Data Lake.

Personal metadata repository that was designed as solution to identified gap in the management of personal data effectively shows that management of metadata can be improved. A similar solution could be applied to other types of metadata than personal, encouraging a vision of graph databases providing a context for data, through describing relations to other data in the data lake.

## 5.7 Applicability of the resulting design on other Data Lake implementations

Although this project was performed on the enterprise architecture of Business Data Lake, a product marketed by Capgemini-Pivotal, the results should be valid for all data lake based architectures.

The abstraction level of designed architecture views was not meant to detail the specific applications used for data processing and data management. Instead, application components were described by their purpose. The advantage of this approach is that solution design is vendor-independent and should be applicable not only on Business Data Lake, but on any solutions for data processing based on data lake architecture model.

In fact, considering that Hadoop and accompanying components and data processing tools are open-sourced technology, it can be argued that the difference between proprietary solutions based on this technology is more that of a flavor than of real substance. Attributing to applicability of the resulting design on other Data Lake implementations is the fact that even the proposed solution design is based on technology that is widely available as open-source.

## 5.8 Applicability of the resulting design on other solutions for data processing

Practical application of a graph database in this project has proved the usefulness of graphs in adding understanding to datasets. Bridging a key gap, graph database was used to map relationships between data subjects and their data, effectively identifying personal data.

Such application possibilities of graph databases for mapping relationships between data are not limited to data lakes. Personal metadata repository is a construct that would solve the problem of personal data management or at the very least improve the performance of other solutions for data processing. The reason behind this assumption is the fact that GDPR requirements are not limited to data lakes, but are set before all solutions for processing of data.

Robinson et al. (2013) clearly states the advantages of graphs over relational tables for describing semantic relationships between data. These advantages are not limited to management of personal data, as they can drastically improve the performance of other areas that rely on relationships between data as well. The reason behind this drastic gain in performance is that queries are localized to only a small portion of the graph regardless of the size of a dataset, eliminating the need for searching the entire dataset, joining tables or other costly queries.

## 5.9 Validation of results

The results of this work are validated through evaluation that is performed by project’s sponsor. This can be judged as reasonable from a business perspective; the investor can be pleased or displeased by contractor’s work. In this case, it is the project sponsor that decides if the resulting design satisfies the requirements, constraints, and principles that were agreed, including the ones related to compliance. This approach is supported by the fact that all these input variables were validated by the project sponsor throughout this work. Moreover, it is clearly in the project sponsor’s best interest that the design actually solves the compliance issues with GDPR, which should motivate a thorough evaluation of the results.

However, a true evaluation of the design can only be done by Data Protection Authority, an EU agency tasked with control over enforcement of GDPR. At the time of this writing, such an agency is yet to be formed.

## 5.10 Validation against the final version of GDPR

At the time of this writing, General Data Protection Regulation is awaiting Council 1st reading position, EU Parliament 1st reading position already being completed. This means that GDPR is in its finalization stage and it is expected to be approved and become law earliest in the second half of 2015 (WSGR, 2015). Under this period, the regulation is reviewed by EU Council and proposals of amendments are being discussed.

The final version of GDPR may include changes of the requirements that the current target architecture solution is not designed to handle, even though such a scenario is not very probable, because the finalization is mostly about details while the designed architecture represents relatively high level of abstraction and is mostly about principles.  Even with that being the case, it is absolutely necessary to validate the results of this work with finalized version of GDPR. Failing to do so, before continuing with the project, would be taking the risk of investing time and resources in a design that is inherently bad. That could lead to ultimate failure of the project and business critical consequences of failing to reach compliance with GDPR.

# 6 Conclusion

The purpose of this chapter is to provide a summary of the performed work, answer the research questions and recommend the future course of the project.

## 6.1 Summary

The primary goal of this thesis was to investigate the compliance of Data Lake architecture model with GDPR and to design the necessary changes in order to make it compliant. This goal can be largely considered as accomplished. The compliance was investigated thoroughly on all three levels of enterprise architecture. The resulting gap analysis effectively revealed the failure of the current architecture model in being compliant with the coming regulation.

The second part of the primary goal was accomplished to the greatest extent possible under the limitations of this thesis. All the changes that need to be made if the compliance is to be achieved are designed on a relatively high level of abstraction. The nature and purpose of these changes is explained through the design of target architectures and the performed gap analysis. A key gap was selected and was bridged through the proposed design of a dedicated repository for personal metadata. The design was based on a graph database that in a revolutionary way improves the performance of personal data management. A prototype was created to demonstrate the design and was made available through code attached to this report.

The secondary goal of this thesis was to shed light on changes that GDPR shall soon enforce upon management of personal data and Big Data in general. This goal can be considered as accomplished, with a reservation that GDPR text may change to some extent during the final stages of its adoption. Through observing the whole design process, which was shown in great detail, the reader can truly understand what changes enforcement of GDPR requires from all participants of data processing operation.

In the process of reaching the goals of this thesis and answering research questions, architectural artifacts were produced. An additional value of this report is that it serves as architecture repository, making artifacts available for supporting the future continuance of development.

In regard to the supporting theory, the results have shown the correctness of Ohlhorst's (2013) theory that the first issue of security and governance in Big Data environment is being able to determine where the data resides. Furthermore, the design of the prototype of a Personal Metadata Repository has demonstrated the advantages of a graph database for storage and management of connected data, confirming the theory of Robinson et al. (2013).

## 6.2 Answers to research questions

**I. Is Data Lake Enterprise Architecture model compliant with GDPR and to what extent?**

The first research question is answered through the performed architecture design and the subsequent gap analysis. The summary of identified gaps is presented in the Table 6 and it shows the existence of a number of gaps on all three layers of enterprise architecture. Some of these gaps are relatively easy to overbridge while the others are more significant. This is reflected on their impact level on achieving compliance with GDPR.

The answer to the first research question is therefore negative; Data Lake model is not compliant with GDPR, at least not in its current state. As to the extent of compliance (or lack of it), gap analysis that was previously performed on all three architectural layers explicitly shows those parts of architecture that are compliant and those that are not. In terms of GDPR requirements, the current – baseline architecture fails to comply on all three levels.

**I. What changes would be necessary to design in order achieve the compliance?**

Table 6 shows the summary of identified gaps, as well as the designed target architecture, providing an explicit answer even for the second research question. As mentioned before, impact level of gaps towards reaching compliance with GDPR differs between gaps. For instance, the capability to create a Request for a copy of personal data is a gap of low impact because of two reasons; for one, to overbridge this gap is not as time and resource consuming as for some other gaps. The other reason is that bridging this particular gap satisfies only a single requirement of GDPR, while bridging some other gaps might benefit the satisfaction of several if not all requirements, causing a much greater impact on overall compliance of the architecture.

By analyzing the target architecture, gaps and gaps impact levels, it can be concluded that there exists one key gap that represents functionality that is used in bridging almost all other gaps. This gap is identified as Personal metadata repository and it enables solution to a number of other gaps by providing the means to locate personal data. This capability is crucial to fulfilling all of the Data subject's requests, Personal data management and in doing so even benefit Unified Data Management. Data subject's request management and Event report for DPA are gaps that will remain unaffected, but it can be argued that these two gaps are not significant issues in the overall compliance problem; Event report for DPA could be solved by a reporting tool and may even be facilitated from the Data Protection Authority's side, while Data subject's request management is an effective way for data subjects to communicate with DPOs, but not an actual GDPR requirement.

Being identified as the gap with highest impact on compliance with GDPR, solution design for Personal metadata repository has been designed.

## 6.3 Critique

Regardless if a project ends successfully or not, there will always be things that could have been done better and things that should be improved in the future. This project was no exception to this rule.

One of the things that I regret the most is not being able to perform a planned interview with the correspondent who is a technical expert on Business Data Lake (BDL in further text), or find a more suitable replacement for that particular area of expertise. This interview was meant to provide a deeper knowledge about the inner workings of BDL that would perhaps have revealed more gaps on the application and technology layers of enterprise architecture. In particular, I had hoped to learn the fine details about file deletion process in BDL and Hadoop in general, that would enable the clarification whether Hadoop immutability issue is manageable or not and what would be the consequences of such an action. Many of the questions were answered through other two interviews and available documentation, but I feel that a higher quality of this work could have been reached by interviewing an expert of this particular area.

Establishing Key Performance Indicators (KPI) for the target design has been problematic, to say the least. The way that KPIs are defined in TOGAF ADM, they should by quantifiable measurements. This was not found to be an optimal solution for indicating the desired state of architecture. How does one measure compliance anyway? By the number of compliance requirements satisfied? Or perhaps by calculating in the individual impact of these requirements on overall compliance as well? I understood this as more of a question of quality than quantity, as partially compliant is still non-compliant. This was a subjective decision and, as such, it could be an error in judgement. Perhaps a deeper analysis of how compliance can be quantified is prudent before the eventual continuance of the design. Developing or using an existing methodology for quantification of compliance would presumably be a better solution for establishing KPIs.

A reasonable amount of critique could be placed at the fact that only one design solution was done in detail. Only one gap, albeit a key one, was bridged through a detailed design. Other gaps have remained on a higher level of abstraction. The reason for such design decision and for concluding the project at this point is that the project is not without limitations. Project's resources and the need for validation of results before continuance are some of them. Another limitation is the academic aspect of this project. Being conducted as a bachelor thesis, this work falls under certain standards, both formal and informal, about its structure, length and complexity. This work already stretches these standards up to a point that further elaboration of target architecture is not prudent, even though the motivation to continue is strong. Although the heart of the solution is formed and a functioning prototype of a Personal metadata repository is created, creating the whole design in greater detail would have to wait until the eventual continuance of the project.

Readiness for change and Business Transformation Risks analysis was done on the base of one interview, available documentation and logical assumptions. Even though these are not meant to be strong points of this work, as change management does not fall into the scope of this project, there is much room for improvement of the way these steps of TOGAF ADM were performed. Relying only on one interview and logical assumptions as sources of knowledge in this matter, apart from available documentation, leaves a huge space for error due to the subjectivity, both from the interviewee and especially from the author of this work.

The previous critique of possible subjectivity is not limited to any particular steps of TOGAF ADM. The scope of this whole project is by no means small and would almost certainly be better served by a larger project team, in which various ideas and visions could flow and be discussed. While ideas were exchanged and discussed with the representative of the project's sponsor, as well with colleagues at the university and even on a seminar dedicated to usability of graph databases, it cannot be certain that something important was not missed on account of subjectivity. In this regard, another set, or several more sets of eyes would have been helpful.

## 6.4 Completeness of this project and proposed work continuance

While this project cannot be regarded as completed in terms of producing an all-encompassing solution to compliance with GDPR, it can be considered completed under the given constrictions and assigned objectives. The objectives were to answer the question if Business Data Lake is compliant with GDPR or not, and to determine the eventual design changes that would be necessary to reach the compliance. These objectives were reached, under given constraints and under the reservation that results would have to be validated against the final version of GDPR.

It must be also noted that GDPR is a truly vast set of regulations that can be transformed into requirements. In this project, a design choice was made to implement those requirements that were considered as key issues in reaching eventual compliance between BDL and GDPR. Other requirements, such as the prohibition of storing and processing personal data of EU citizens outside the EU, were not implemented at this point. The reason for that decision was that such requirements do not cause significant design changes of Business Data Lake's architecture. Nevertheless, it would be a mistake to believe that the difficulty of requirement's implementation is directly proportionate to requirement's importance to compliance with GDPR. Eventual continuance of the project would have to successively broaden the spectrum of GDPR requirements that are addressed. To make a selection and start with the harder ones is always a good idea, because less time and resources would be spent on the project in the case of coming across an unbridgeable gap, or across a gap whose bridging would influence solutions to other requirements.

A step towards the future development was taken by designing the solution for personal metadata repository, as a demonstration of how the work should continue. That what should follow is designing detailed solutions for bridging all identified gaps and creating a prototype in order to test the solutions.

Working on the prototype of personal metadata repository, it became apparent that the use of graph database for managing metadata has a much broader potential than management of personal data. Using a graph for management of data in general is accomplished as simple as adding new properties to nodes of an existing prototype. This could be very useful for improvement of business analytics and should be researched further.

Continuing to a detailed design would cause the workload to grow exponentially, craving for corresponding growth of resources invested in the project.

A prudent choice would be to wait until the final version of GDPR and the validation of this project's results before eventual continuance of the project. At that time, a final confirmation would be gained if the results make a good base for future development or not. Furthermore, a total list of GDPR requirements could be constructed at that point, which is something that was not considered feasible during the course of the project so far, due to the fact that the list would not be definite and subject to certain change until GDPR is finalized.

However, my recommendation for the absolutely first step in this project's eventual continuance would be solving the Hadoop immutability issue that has previously been described (see 5.4 Immutability of HDFS). This is a critical issue in regard to Data Lake's compliance with the GDPR that was not mitigated during the course of this work and is therefore the logical choice for future development.

## 6.5 Reflection

The paradigm of design research is gaining knowledge about problems and their solutions through design. This can even open new problems that must be tackled if the original problem is to be solved. This thesis was no exception and many new discussion topics were opened.

So, when can one design be considered finalized? At what point does the continuance of the design work stop to produce new knowledge? The resulting design is a significant step forward towards reaching an optimal solution to the problem, but it also opened new questions and avenues of research. Nevertheless, the overwhelming scope of the problem must be considered when making a reflection on this thesis.

At the very beginning of the thesis, the writer of these lines was enthusiastic. A challenging task was received from a prestigious company in the form of an extremely current and tangible problem, a solution to which held a promise of a great impact both to management of personal information in the Big Data environment and to the professional development of the author. Then, as the tedious and grinding work of developing an Enterprise Architecture single-handedly began, the future appeared less bright. There was no end to business scenarios, diagrams, requirements and architecture views. It was easy to take a lesser path, give up, perhaps take on a less significant problem and do just enough to get over the finish line. The choice was made to persevere and grind the problem right back. In the end, the research questions were answered and the design of a solution presented.

The first sign of light at the end of the tunnel came during gap analysis, when the contours of the solution started to take shape. By gaining new energy through this insight, the work was completed and a proposition was made for a first step of the solution to unavoidable compliance issues that Business Data Lake architecture model and Big Data management in general will have with GDPR.

# 7 Bibliography

- Bennet, S., et al., (2010), Object-Oriented Systems Analysis and Design, McGraw Hill
- Brewer, Brian, (2015), Big Data Must Have Metadata, InfoLibrarian – Best Practices Blog, Retrieved 2015-05-06 from [http://www.infolibcorp.com/blog/big-data-metadata/big-data-must-have-metadata/]
- Capgemini, (2013), The Technology of the Business Data Lake, Capgemini, Retrieved 2015-05-06 from [http://www.capgemini.com/resource-file-access/resource/pdf/pivotal-business-data-lake-technical_brochure_web.pdf]
- Chaudhuri, et al., (2011), An Overview of Business Intelligence Technology, article published in Communications of the ACM, August 2011, vol. 54, no.8, Retrieved 2015-05-06 from [http://cacm.acm.org/magazines/2011/8/114953-an-overview-of-business-intelligence-technology/fulltext]
- Chen, Hsinchun, et al., (2012), Business Intelligence and Analytics: From Big Data to Big Impact, article published in MIS Quarterly Vol. 36 No. 4, pp. 1165-1188/December 2012, Retrieved 2015-05-06 from [http://ai.arizona.edu/mis510/other/MISQ%20BI%20Special%20Issue%20Introduction%20Chen-Chiang-Storey%20December%202012.pdf]
- CITO Research, 2014, Putting the Data Lake to Work: A Guide to Best Practices, CITO Research whitepaper sponsored by Teradata and Hortonworks, Retrieved 2015-05-06 from [http://hortonworks.com/wp-content/uploads/2014/05/TeradataHortonworks_Datalake_White-Paper_20140410.pdf]
- Davis, Ben, (2014), Metadata Management in the Data Lake, Teradata International Blog, Retrieved 2015-05-06 from [http://blogs.teradata.com/international/metadata-management-in-the-data-lake/]
- Davis, Mike, (2014), Making sense of European Data Protection Regulations as they relate to the storage and management of content in the Cloud, AIIM, Retrieved 2015-05-06 from [www.aiim.org]
- EMC, (2014), Security and Compliance for Scale-out Hadoop Data Lakes, EMC whitepaper, Retrieved 2015-05-06 from [http://www.emc.com/collateral/white-paper/h13354-wp-security-compliance-scale-out-hadoop-data-lakes.pdf]
- European Digital Rights, (2014), Key aspects of the proposed General Data Protection Regulation explained, Retrieved 2015-05-06 from [www.edri.org]
- European Commission, (2014), Progress on EU data protection reform now irreversible following European Parliament vote, Retrieved 2015-05-06 from [http://europa.eu/rapid/press-release_MEMO-14-186_en.htm]
- European Commission, (2015), Data Protection Day 2015: Concluding the EU Data Protection Reform essential for the Digital Single Market, European Commission - Fact Sheet, Retrieved 2015-05-06 from [http://europa.eu/rapid/press-release_MEMO-15-3802_en.htm]
- European Parliament, (2014), European Parliament legislative resolution of 12 March 2014 on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (COM(2012)0011 – C7-0025/2012 – 2012/0011(COD)), Retrieved 2015-05-06 from

[http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P7-TA-2014-0212+0+DOC+XML+V0//EN]

- Harper, Jelani, (2014), Data Lakes: Complicating Big Data Governance, article published on Dataversity 2014-11-06, Retrieved 2015-05-06 from [http://www.dataversity.net/data-lakes-complicating-big-data-governance/]
- Hevner, A. R., (2004), Design science in information systems research, MIS quarterly. 28(1): p75–105, Retrieved 2015-05-06 from [http://em.wtu.edu.cn/mis/jxkz/sjkx.pdf]
- Hurtado, Joseph, (2014), Open Kanban: The First Agile and Lean Open Source Method for Continuous Improvement, InfoQueue, Retrieved 2015-05-06 from [http://www.infoq.com/articles/open-kanban-introduction]
- Ipswitch, (2014), European IT Teams Woeful Lack of Preparation for General Data Protection Regulation (GDPR) May Mean Painful Compliance Audits Ahead, Retrieved 2015-05-06 from [http://www.ipswitchft.com/about-us/news/press-releases/2014/11/gdpr-may-mean-painful-compliance-audits-ahead]
- Jacobs, Adam, (2009), Pathologies of Big Data, article published in Communications of the ACM, August 2009, vol. 52, no.8, Retrieved 2015-05-06 from [http://dl.acm.org/citation.cfm?id=1536632]
- Kuner, Christopher, et al., (2012), The challenge of 'big data' for data protection, article published in Oxford Journals, International Data Privacy Law, volume 2, Issue 2, p.47-49, Retrieved 2015-05-06 from [http://idpl.oxfordjournals.org/content/2/2/47.full]
- Lankhorst, M. et al., (2013), Enterprise Architecture at Work, Springer
- Lu, Ruopeng et al., (2008), Compliance Aware Business Process Design, article published in Business Process Management Workshops, Lecture Notes in Computer Science Volume 4928, 2008, pp 120-131
- Maier, Andrew, (2010), Complete Beginner's Guide to Design Research, article published on UX Booth, Retrieved 2015-05-06 from [http://www.uxbooth.com/articles/complete-beginners-guide-to-design-research/]
- Maxwell, C. J. and Anton, I. A., (2010), The production rule framework: developing a canonical set of software requirements for compliance with law, article published in IHI '10 Proceedings of the 1st ACM International Health Informatics Symposium, p. 629-636, ACM New York, Retrieved 2015-05-06 from [http://dl.acm.org/citation.cfm?id=1883092]
- Narayanan, A., Shmatikov, V., (2008), Robust De-anonymization of Large Sparse Datasets, University of Texas at Austin, Retrieved 2015-05-06 from [https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf]
- Ohlhorst, Frank, (2013), Big Data Analytics – Turning Big Data into Big Money, Wiley & Sons
- Ohm, P., (2010), Broken promises of privacy: Responding to the surprising failure of anonymization, UCLA Law Review, vol. 57, 2010, pp. 1701, Retrieved 2015-05-06 from [http://uclalawreview.org/pdf/57-6-3.pdf]
- Oliver, C., Andrew, (2014), How to create a data lake for fun and profit, InfoWorld, Retrieved 2015-05-06 from [http://www.infoworld.com/article/2608490/application-development/how-to-create-a-data-lake-for-fun-and-profit.html]
- Oracle, (2015), An Enterprise Architect's Guide to Big Data: Reference Architecture Overview, Oracle Enterprise Architecture Whitepaper, February 2015, Retrieved 2015-05-06 from

[http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf]

- Robinson, I., Webber, J., Eifrem, E., (2013), Graph Databases, O'Reilly Media
- Sawant, N., Shah, S., (2013), Big Data Application Architecture Q & A, Apress Media
- Shaw, J., (2014), Why Big Data Is a Big Deal, article published in Harvard Magazine March-April 2014, Retrieved 2015-05-06 from [http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal]
- Stein, Brian and Morrison Alan, (2014), The enterprise data lake: Better integration and deeper analytics, PWC Technology Forecast, Retrieved 2015-05-06 from [http://www.pwc.com/en_US/us/health-industries/assets/pwc-tech-forecast-data-lakes.pdf]
- Tene, Omer., Polonetsky, Jules, (2012), Privacy in the Age of Big Data - A Time for Big Decisions, Stanford Law Review Online, Retrieved 2015-05-06 from [http://www.stanfordlawreview.org/online/privacy-paradox/big-data]
- The Open Group, (2011), TOGAF Version 9.1, The Open Group, Retrieved 2015-05-06 from [http://pubs.opengroup.org/architecture/togaf9-doc/arch/]
- The Open Group, (2015), Membership report, The Open Group, Retrieved 2015-05-06 from [http://reports.opengroup.org/membership_report_all.pdf]
- The Open Group, (2015), Archimate, The Open Group, Retrieved 2015-05-06 from [http://www.opengroup.org/subjectareas/enterprise/archimate]
- Triger, Loic, (2014), Changes in European Data Protection Regulation: A look at the GDPR, article published on Techradar 2014-12-30, Retrieved 2015-05-06 from [http://www.techradar.com/news/internet/policies-protocols/changes-in-european-data-protection-regulation-a-look-at-the-gdpr-1278235]
- Villion, Matt, (2012), Security Think Tank: How to prepare for EU data protection rules (part 7), article published on Computer Weekly, Retrieved 2015-05-06 from [http://www.computerweekly.com/opinion/Proposed-EU-Data-Protection-Regulation-what-should-companies-be-thinking-about]
- White, A., Heudecker, N., (2014), The Data Lake Fallacy: All Water and Little Substance, Gartner research report, Retrieved 2015-05-06 from [https://www.gartner.com/doc/2805917/data-lake-fallacy-water-little]
- White, Tom, (2012), Hadoop: The Definitive Guide, third edition, O'Reilly
- WSGR, (2015), EU legislative process updates , Wilson, Sonsini, Goodrich & Rosati LLP, Retrieved 2015-05-06 from [https://www.wsgr.com/eudataregulation/process-updates.htm]
- Zikopoulos et al., P., et al.,(2012), Harness the Power of Big Data – The IBM Data Platform, McGraw-Hill

# Appendix A

## Time schedule of interviews

| Date | Correspondent | | Interview subject |
|---|---|---|---|
| | Name | Position | |
| 2015-02-15 Kick-off meeting | Christofer Holmgren Bagge<br>Role: Stakeholder | Managing Consultant and Senior Business Analyst with focus on Big Data and Big Data Analytics at Capgemini Sweden | Problem background, Problem description, Business drivers, Requirements, GDPR's impact on Big Data processing, Baseline, Available system documentation |
| Interview failed on account of the unavailability of the intended correspondent. The information is instead procured through available documentation and consultation with project sponsor. | Tobias Karlsson<br>Role: Stakeholder/External Data Lake expert | Senior Field-engineer at Pivotal Sweden | Business Data Lake, Data ingestion, Metadata generation, Unified Management Tier |
| 2015-03-17 | Neo4j Graph Day Stockholm, Neo4j lectures and<br>1 hour one on one interview with Jim Webber, Ph. d., Chief Scientist at Neo4j<br>Role: Graph DB expert | Neo4j graph database Chief scientist | Neo4j under Hadoop, using graph databases for cataloging data ownership, linking data with owner metadata with Neo4j in Data Lake |
| 2015-05-07 | Christofer Holmgren Bagge<br>Role: Stakeholder | Managing Consultant and Senior Business Analyst with focus on Big Data and Big Data Analytics at Capgemini Sweden | Evaluation |

Table A1, Interview schedule

## Interview questions for semi-structured interviews

| Date | Correspondent | | Interview questions |
|---|---|---|---|
| | Name | Position | |
| 2015-02-15 Kick-off meeting | Christofer Holmgren Bagge<br>Role: Stakeholder | Managing Consultant and Senior Business Analyst with focus on Big Data and Big Data Analytics at Capgemini Sweden | • Can you please describe the Business Data Lake (BDL)?<br>• Can you describe the problem background?<br>• Can you describe the concrete problem?<br>• Which are the stakeholders to BDL?<br>• Which are the stakeholders to the compliance issue?<br>• How would you describe business drivers, concerns and requirements of Business Data Lake?<br>• How would you describe business drivers, concerns and requirements of Business Data Lake connected to compliance with GDPR?<br>• In your opinion, what are the key changes that GDPR would cause to the way Big Data is processed?<br>• How critical is compliance with GDPR to Big Data processing?<br>• How do you expect that GDPR will be enforced in practice, what are the key obstacles in regard to compliance?<br>• Could you please describe how the system is functioning today in regard to compliance?<br>• How is personal data recognized and managed today?<br>• Is there a way to extract metadata about the data owner at ingestion stage?<br>• What impact would that cause to system's performance?<br>• What would, in your opinion, be an optimal solution to compliance issue between BDL and GDPR?<br>• Where can I find system documentation? |
| Interview failed on | Tobias Karlsson<br>Role: | Senior Field-engineer at | • Could you please describe how does BDL work exactly? |

|  |  |  |  |
|---|---|---|---|
| account of the unavailability of the intended correspondent. The information is instead procured through available documentation and consultation with project sponsor. | Stakeholder/External Data Lake expert | Pivotal Sweden | • How is personal data managed inside BDL today?<br>• How is data generally managed in BDL?<br>• Could you please describe the Data ingestion process in the BDL?<br>• How is metadata extracted and stored during this process?<br>• What are the advantages and downsides of systematic metadata extraction at data ingestion stage?<br>• Could you please describe the operation of the Unified Management Tier in detail?<br>• Where can I find system documentation? |
| 2015-03-17 | Neo4j Graph Day Stockholm, Neo4j lectures and<br>1 hour one on one interview with Jim Webber, Ph. d., Chief Scientist at Neo4j<br>Role: Graph DB expert | Neo4j graph database Chief scientist | • Are you acquainted with GDPR?<br>• Are you acquainted with BDL?<br>• What would, in your opinion, be the most optimal solution for cataloging data ownership in Big Data ecosystem and why?<br>• How do you propose that this solution would be implemented under HDFS2 file system?<br>• How to solve versioning issue – existence of multiple instances of information across Hadoop?<br>• What advantages could be gained by using graph database to link the data with its owner?<br>• What challenges would this solution present?<br>• How many Neo4j server instances are required to service a Data Lake in a Big Data environment?<br>• Where can I find system documentation? |
| 2015-05-07 | Christofer Holmgren Bagge<br>Role: Stakeholder | Managing Consultant and Senior Business Analyst with focus on Big Data and Big Data Analytics at Capgemini Sweden | Would you please provide an evaluation and comments on the project, with focus on method and accomplished results? |

Table A2, Interview questions

# Appendix B

## Description of Business Data Lake

The Figure B1 depicts the key tiers of Business Data Lake Architecture, with data flowing from left to right. Velocity aspect of data is presented vertically, with data at rest concentrated on lower levels, while real-time transactional data is shown on higher levels of the Figure B1.

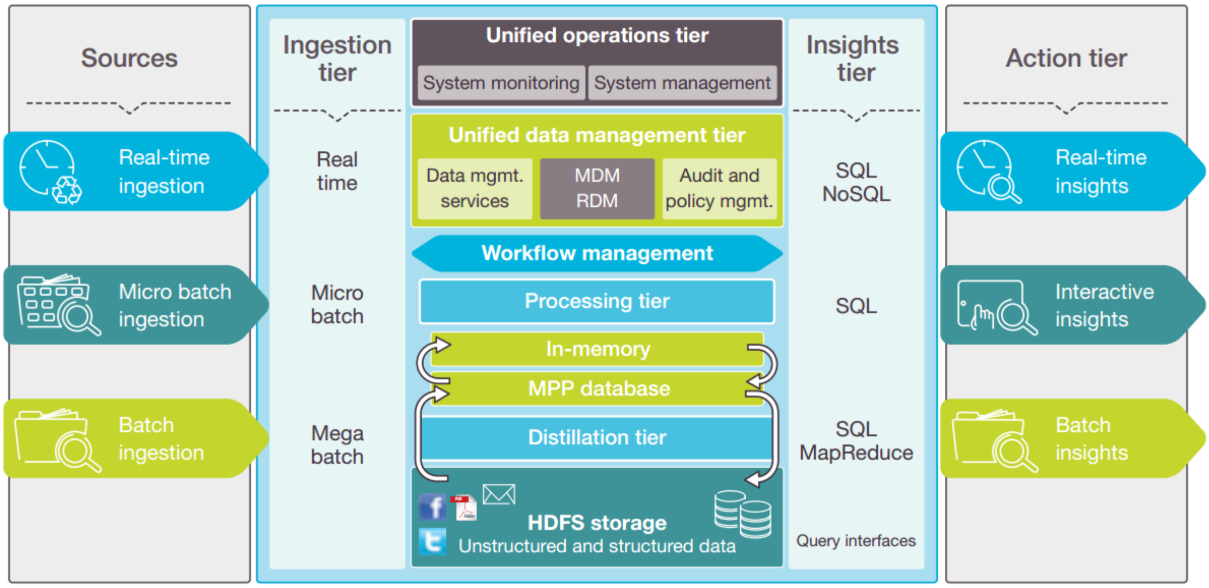


Figure B1, Tiers of Business Data Lake (Capgemini 2013)

Data storage tier provides the storage of all the data plus real-time access to selected data. The Variety and Volume aspects of Big Data created the demand for a cost-effective, reliable storage mechanism which came in the form of the Hadoop Distributed File System (HDFS2). This solution provides a low-cost storage for all the data at rest in the system. Data is stored in its original form, meaning that the necessary structure is added as needed, using “schema-on-read” principle and avoiding the need for heavy extract transform load (ETL) processing of data as it is ingested. The need for real-time responses on some data presents difficulties to HDFS2 as the latency of writing the data to disk introduces too much delay. This problem is solved by using an in-memory data solution (Gemfire in-memory database) for handling real-time data in motion, while data at rest is stored on HDFS2.

Distillation tier provides structure to the stored, raw data when needed, so that analytical or machine learning algorithms can be employed and supporting streaming of individual events or transactions into the in-memory portion of the data lake where they can be processed with very low latency.

Unified data management tier serves to manage and provide access to all the data in the Business Data Lake. Authorized data workers can access data sets through a self-service portal that enables them access to metadata catalog of the data in the system and to create a single view of data from across the enterprise. The access is regulated by policy-controlled leases, with the resources being released back to the system upon lease expiration. The Figure B2 depicts the architecture of the unified data management tier.

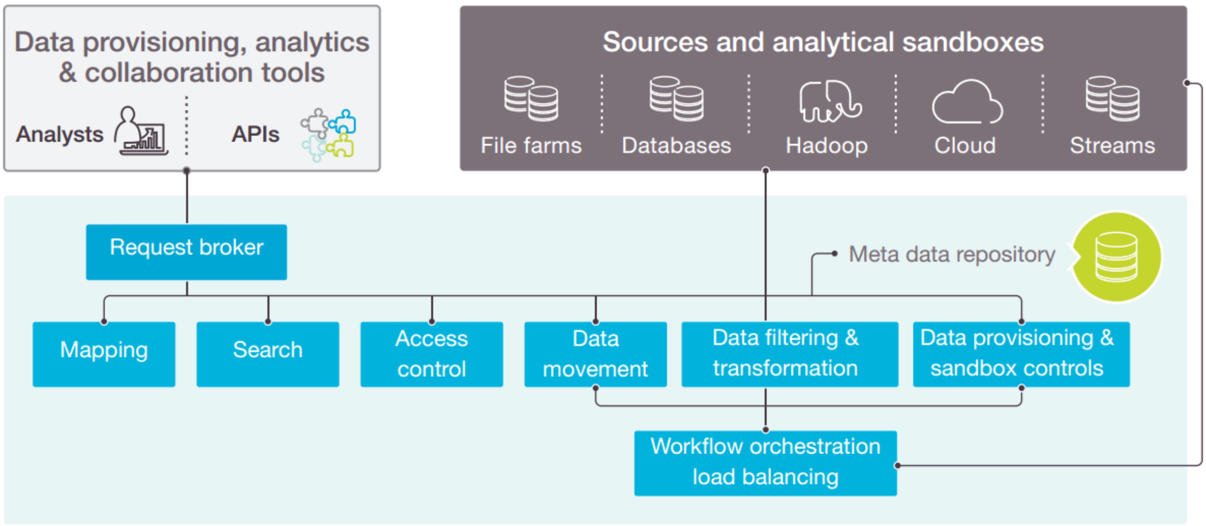

Data provisioning, analytics & collaboration tools
Analysts
APIs
Sources and analytical sandboxes
File farms
Databases
Hadoop
Cloud
Streams
Request broker
Meta data repository
Mapping
Search
Access control
Data movement
Data filtering & transformation
Data provisioning & sandbox controls
Workflow orchestration load balancing

Figure B2, Unified Management Tier (Capgemini, 2013)

Insights tier of the Business Data Lake supports access through a variety of Hadoop query interfaces such as Hive or Pig, but also through traditional SQL query. Interactive analytics tools and graphical dashboards enable joining of data across different data sets in order to draw insights. Insights tier also supports usage of MADlib analytic libraries for performing mathematical, statistical, and machine learning analysis, as well as real-time insights that can be set to generate external events. (Capgemini, 2013)

# Appendix C

## Architecture Principles

### Business Principles

| Name | Primacy of Principles |
|---|---|
| Reference | BP01 |
| Statement | These Architecture Principles apply to all organizations within the enterprise. |
| Rationale | Consistent and measurable quality of information supplied to decision-making process can only be guaranteed if all enterprise organizations uniformly accept and apply the established principles. |
| Implications | All information management initiatives must be compliant with the principles. If an initiative stands in conflict with a principle, the framework of the initiative must be modified in order to comply with the principle. |

Table C1, BUSINESS PRINCIPLE 01

| Name | Requirements-Based Change |
|---|---|
| Reference | BP02 |
| Statement | Changes to applications and technology are made only in response to business needs. |
| Rationale | Information environment is changed in response to the needs of business, not the other way around. |
| Implications | No technical improvement or system development is made unless a documented business need exists. Responsive change principle and keeping the system up to date is also a business requirement. |

Table C2, BUSINESS PRINCIPLE 02

| Name | Information Management serves Enterprise |
|---|---|
| Reference | BP03 |
| Statement | To provide maximal benefit to the enterprise as a whole is the reason behind information management decisions. |
| Rationale | Maximum benefit to the enterprise is the main business driver and therefore the most important reason for information management decisions to be made. |
| Implications | Information management decisions made throughout the enterprise must be aligned with the enterprise plan. |

Table C3, BUSINESS PRINCIPLE 03

| Name | Business Continuity |
|---|---|
| Reference | BP04 |
| Statement | Continuity of business processes must be ensured. |
| Rationale | Business processes are dependent on underlying architectures. Therefore, the reliability of these architectures must be ensured throughout their design and use. Events that can compromise business continuity are hardware failure, natural disasters, and data corruption, but also intentional or unintentional security breach. The risk of such events must be mitigated in order to prevent the disruption of enterprise activities. |
| Implications | Risks of business interruption must be identified and managed. Continuity of business-critical functions must be ensured through redundant or alternative capabilities. Business continuity solution should be part of the design from the start. |

Table C4, BUSINESS PRINCIPLE 04

| Name | Service-Oriented Architecture |
|---|---|
| Reference | BP05 |

| | |
|---|---|
| **Statement** | The architecture is based on services that mirror business activities. |
| **Rationale** | Service orientation makes it easier to map business activities to services that help carry them out. This way it is plain to see which service corresponds to which business activity and subsequently decide which service needs to be improved based on business performance. |
| **Implications** | Business descriptions must be provided in order to give context to the services. Governance of services is required to ensure that a service always have an actual business activity it carries out. |

Table C5, BUSINESS PRINCIPLE 05

| **Name** | **Compliance with Law** |
|---|---|
| **Reference** | BP06 |
| **Statement** | Enterprise Architecture must facilitate compliance with all relevant laws, policies and regulations. |
| **Rationale** | Non-compliance can lead to both material and criminal liability and business-critical consequences to both users and suppliers of the Enterprise Architecture. |
| **Implications** | The enterprise must take steps to comply with laws, regulations, and external policies regarding the collection, retention, and management of data. Compliance is a very important business driver. |

Table C6, BUSINESS PRINCIPLE 06

| **Name** | **IT Organization’s Responsibility** |
|---|---|
| **Reference** | BP07 |
| **Statement** | The IT organization’s responsibility is to own and implement IT projects that result in solutions to business needs that meet user-defined requirements for functionality, service levels, cost, and delivery. |
| **Rationale** | All IT projects should be performed in accordance to assigned resources and given time-frame. Expectations must be aligned with capabilities in order for projects to be cost-effective. |
| **Implications** | Projects must be prioritized according to their benefit to the enterprise. Integrated solutions should be enabled by data, application, and technology models. |

Table C7, BUSINESS PRINCIPLE 07

| **Name** | **Intellectual Property Protection** |
|---|---|
| **Reference** | BP08 |
| **Statement** | Intellectual property of the enterprise must be protected. |
| **Rationale** | As the primary function of the enterprise is to divine actionable information from collected data, it is clear that both collected data and the divined information represent competitive advantage, have a measurable value, and is therefore classified as Intellectual property that must be protected. |
| **Implications** | Much of the actual protection of intellectual property is implemented in the IT domain. Security policy is required. |

Table C8, BUSINESS PRINCIPLE 08

### Data Principles

| Name | Data is an Asset |
|---|---|
| Reference | DP01 |
| Statement | Data is an asset that has value to the enterprise and is managed accordingly. |
| Rationale | Data is the foundation of our decision-making and is a valuable corporate resource. As such, it has a real and measurable value. |
| Implications | Responsibilities must be assigned at an enterprise level to managing data. |

Table C9, DATA PRINCIPLE 01

| Name | Data is Shared |
|---|---|
| Reference | DP02 |
| Statement | Data is shared across enterprise functions and organizations. |
| Rationale | Using only one source of data across the enterprise creates data consistency and leads to more unison decision-making because every enterprise unit makes its decisions based on the very same data as any other unit. The management also has a better insight into which data are decisions of subordinate units based on. |
| Implications | Data Lake is the enterprise-wide "virtual single source" of data. A general set of policies, procedures and standards governing data management and access must be developed in order for data to be shared on enterprise level. Therefore an Enterprise Architecture must enable implementation of said policies, procedures and standards. It is important to develop metadata that defines this shared environment and develop a repository system for storing this metadata to make it accessible. A care must be taken that data sharing principle does not jeopardize the principle of data security. |

Table C10, DATA PRINCIPLE 02

| Name | Data is Accessible |
|---|---|
| Reference | DP03 |
| Statement | Data is accessible to users throughout the enterprise, but only through access control. |
| Rationale | Wide access to data leads to efficiency and effectiveness in decision making. |
| Implications | Business Data Lake model enables users to access the data that they need, using the tools necessary to perform their task. Access to data does not necessarily grant the user access rights to modify or disclose the data. A strong access control must also be established. |

Table C11, DATA PRINCIPLE 03

| Name | Data is Usable |
|---|---|
| Reference | DP04 |
| Statement | Data must be usable to a variety of enterprise functions and organizations across the enterprise. |
| Rationale | Transforming the data before sharing it, constrains the use of data because when data is cleaned and structured for use by one enterprise unit, it can lose the properties of importance to another. The value of data must be preserved regardless of intended purpose. |
| Implications | Data must be ingested and stored in its original form in order to remain usable for all purposes. Instead of structuring data on Write, when it is ingested, data is structured on Read, based on user's query. Unstructured data may present governance and compliance issues. Data access control may also present a challenge because without knowing what the data actually contains, one cannot assign the appropriate access rights. |

Table C12, DATA PRINCIPLE 04

| Name | Data is personal |
|---|---|
| Reference | DP05 |
| Statement | Personal data must be managed in accordance with relevant legislation. |
| Rationale | Much of the data that is ingested, stored and used contains information that classifies as personal data and must be managed according to specific rules. This principle becomes even more important with the upcoming GDPR legislation that broadens the definition of personal data. |
| Implications | Already from ingestion stage, personal data must be identified in order for it to be managed. In order to satisfy requirements for managing personal data, a register of this data must be maintained. A responsibility for managing personal data must be implemented on enterprise level. |

Table C13, DATA PRINCIPLE 05

| Name | Data is secure |
|---|---|
| Reference | DP06 |
| Statement | Data must be protected from unauthorized use and disclosure. |
| Rationale | Unauthorized use and disclosure of data may lead to business-critical consequences. This is true both in the case of data that is classified as important for business continuity and in the case of personal data. |
| Implications | Usability of data must be balanced against the need for protection. Security classification of data must be implemented. Aggregation of data may result in an increased classification level. Security should be a part of the design process already from the start; it is much more difficult to add it later. All levels of EA must be protected from unauthorized access and manipulation. Security not only includes inadvertent or unauthorized alteration and disclosure, but also sabotage or a disaster. |

Table C14, DATA PRINCIPLE 06

| Name | Data is big |
|---|---|
| Reference | DP07 |
| Statement | Big Data comes in large volume, at great variety and great velocity. |
| Rationale | Managing and analyzing Big Data has become a great challenge to enterprises. The data comes in extreme volumes, which are alone resource-challenging to manage and analyze. Additionally, it comes in great variety of data types, both structured and unstructured.  The data also comes at different speeds and requires different speed of analysis. The most challenging situation thinkable is real-time analysis of Big Data from real-time data streams. |
| Implications | Big Data presents a challenge on an enterprise level that cannot be managed by traditional solutions, such as Data Warehouses. A new solution, specifically designed for managing Big Data is required. |

Table C15, DATA PRINCIPLE 07

| Name | Common Vocabulary of Data Definitions |
|---|---|
| Reference | DP08 |
| Statement | Data must be defined in the same way throughout the enterprise. |
| Rationale | Because data is shared throughout the enterprise, a common vocabulary and data definitions must exist in order for communication between enterprise units to be effective. |
| Implications | Data definitions must be consistent, understandable and communicated to all users through appropriate user training. Data administration responsibilities must be assigned. Ambiguities of data definitions must be eliminated. |

Table C16, DATA PRINCIPLE 08

### Application Principles

| Name | Technology Independence |
|---|---|
| Reference | AP01 |
| Statement | Applications should, whenever possible, be independent of specific technology platforms. |
| Rationale | Technology continually gets obsolete. In order to avoid the collision with BP02, which is to avoid that technology becomes a business driver, applications should not be dependent on specific hardware and operating systems software. Furthermore, this principle promotes a more cost-effective way for applications to be developed, upgraded, and operated. |
| Implications | Portability of applications should be considered. A middleware should be used to decouple applications from underlying software and hardware. Using Java applications is a viable alternative for accomplishing this. |

Table C17, APPLICATION PRINCIPLE 01

| Name | Ease of use |
|---|---|
| Reference | AP02 |
| Statement | Applications must be easy to use by normal users. |
| Rationale | The easier an application is to use, the more productive the user is. Other advantages are that training is kept to a minimum, as it is proportional to application's difficulty grade and the risk for error while using a system is lower. |
| Implications | Integrated information environment is preferable to using isolated systems because it provides a common look and feel to different tasks. |

Table C18, APPLICATION PRINCIPLE 02

| Name | Common Use Applications |
|---|---|
| Reference | AP03 |
| Statement | Applications that can be used commonly across the enterprise are preferable to similar or duplicative applications. |
| Rationale | Different applications that have similar functionality represent unnecessary cost and can produce inconsistent and conflicting data. |
| Implications | Enterprise units should not be allowed to develop applications that are similar in function to enterprise-wide applications. Data sharing principle depends on consistency of data and therefore must all enterprise units' capabilities that produce different data be replaced with enterprise-wide capabilities that produce the single instance of data. |

Table C19, APPLICATION PRINCIPLE 03

### Technology Principles

| Name | Interoperability |
|---|---|
| Reference | TP01 |
| Statement | Technology should conform to defined standards that promote interoperability for data, applications, and technology itself. |
| Rationale | Conforming to standards helps inoperability, which in turn improves the ability to manage systems, user satisfaction, and maximizes ROI by reducing costs. |
| Implications | Interoperability standards and industry standards, preferably open sourced, will be followed unless there is a compelling business reason to implement a non-standard solution. Standards that are to be followed must first be defined and communicated throughout the enterprise. |

Table C20, TECHNOLOGY PRINCIPLE 01

| **Name** | **Scale** |
|---|---|
| **Reference** | TP02 |
| **Statement** | Technology should support scalability. |
| **Rationale** | Explosive increase of Big Data is a well-documented trend. Growth by scale is necessary in order to keep up with this trend. Economy of scale is much more preferable than the economy of developing new systems. |
| **Implications** | Growth by scale must be supported on all levels of enterprise architecture. Scalability means parallelization and the use of common instead of special hardware. Scalability will also enable economic growth or economic stagnation in response to business needs of the enterprise. |

Table C21, TECHNOLOGY PRINCIPLE 02

| **Name** | **Control Technological Diversity** |
|---|---|
| **Reference** | TP03 |
| **Statement** | Technological diversity must be kept to a minimum, unless there is a legitimate business need to do otherwise. |
| **Rationale** | Supporting alternative technologies for processing environments presents a non-trivial cost. Using common technology throughout the enterprise also makes scaling more economical. |
| **Implications** | Interoperability standards and industry standards, preferably open sourced, will be followed unless there is a compelling business reason to do otherwise. Standards that are to be followed must first be defined and communicated throughout the enterprise. A responsibility should be assigned for standardization. |

Table C22, TECHNOLOGY PRINCIPLE 03

# Appendix D

## Request for Architecture Work

### Document Information

| | | | |
|---|---|---|---|
| **Project Name:** | Data Lake Compliance 2015 | | |
| **Prepared By:** | Vuk Kadenic | **Document Version No:** | 0.1 |
| **Title:** | Request for Architecture Work | **Document Version Date:** | 15-02-26 |
| **Reviewed By:** | | **Review Date:** | |

### Distribution List

| **From** | **Date** | **Phone/Fax/Email** |
|---|---|---|
| Vuk Kadenic | 15-02-26 | vukkad-1@student.ltu.se |
| | | |

| **To** | **Action*** | **Due Date** | **Phone/Fax/Email** |
|---|---|---|---|
| Christofer Holmgren Bagge | Approve | 15-03-02 | ******************* |
| LTU | Review | 15-03-02 | |
| | | | |
| | | | |

* Action Types: Approve, Review, Inform, File, Action Required, Attend Meeting, Other (please specify)

### Document Version History

| **Version Number** | **Version Date** | **Revised By** | **Description** | **Filename** |
|---|---|---|---|---|
| 0.1 | 15-02-26 | Vuk Kadenic | Request for Architecture Work, first version. | Request for Architecture Work.docx |
| | | | | |

# Purpose of this Document

This document is a Request for Architecture Work for the Data Lake Compliance 2015 project.

A Request for Architecture work describes the business imperatives behind the architecture work, thus driving the requirements and performance metrics for the architecture work. This should be sufficiently clear so that initial work may be undertaken to scope the business outcomes and resource requirements, and define the outline information requirements and associated strategies of the architecture work to be done.

The Request for Architecture Work is a document that is sent from the sponsoring organization to the architecture organization to trigger the start of an architecture development cycle. Requests for Architecture Work can be created as an output of the Preliminary Phase, a result of approved architecture Change Requests, or terms of reference for architecture work originating from migration planning.

In general, all the information in this document should be at a high level.

# Request for Architecture Work

## Summary of Request

This request concerns investigating the compliance of the Data Lake model with GDPR (General Data Protection Regulation) and designing the necessary changes in order to make it compliant. As a baseline architecture model, Capgemini-Pivotal's Business Data Lake is to be used.

## Organization Sponsors

This architecture work is requested and sponsored by:

- <<Name>> Christofer Holmgren Bagge
- <<Position>> Managing Consultant and Senior Business Analyst with focus on Big Data and Big Data Analytics
- <<Organization>> Capgemini Sweden
- <<Email>>***************
- <<Tel>>*****************

# Business Imperative

## Organization Mission Statement

Capgemini creates and delivers business and technology solutions that fit the user's needs and drives the results that the user wants.

## Business Goals (and Changes)

Business goals are:

- Improve Profitability
- Increase Market Share
- Improve Company Image
- Improve Business Process Performance
- Improve Effectiveness of IT Organization
- Improve User Productivity
- Improve Portability and Scalability
- Improve Interoperability
- Decrease Vendor Dependency
- Improve Security
- Improve Compliance

## Strategic Plans of the Business

Strategic plan of relevance to this project is to make Business Data Lake a leading solution to managing and deriving Business Intelligence from Big Data.

## Changes in the Business Environment

The change in the business environment that is the most immediate business driver for this architecture work is the introduction of new legislation – the GDPR (General Data Protection Regulation). GDPR presents new, challenging requirements in the area of compliance that must be met in order for Business Data Lake architecture model to remain one of the means toward reaching the Business Goals.

## Purpose of Architecture Work

The purpose of this Architecture Work is to investigate the compliance of the Data Lake model with GDPR (General Data Protection Regulation) and to design the necessary changes in order to make it compliant. Because Capgemini-Pivotal's Business Data Lake will be used as baseline architecture model, the investigation and the subsequent changes would directly help Capgemini in overcoming the compliance challenge placed before their Business Data Lake Product. The importance of this task is not only relevant to parts of Capgemini's organization that are involved with Business Data Lake as a product, but also to Capgemini as a whole because proposed penalties for non-compliance go up to 5% of global turnover for the whole organization, limited to 100M Euro.

## Success Criteria

An ideal outcome of this project would be Enterprise Architecture model of Business Data Lake that is compliant to GDPR, while its functionality remains the same or better.

An outcome where the architecture model is compliant to GDPR, but the functionality has decreased to a level where it still fulfills its business function of being able to manage Big Data and produce actionable information out of it, in a satisfactory way, would also be considered a success.

Outcomes where the architecture model's compliance is not achieved or where functionality has decreased to such a degree that the architecture is unable to fulfill its business function in a satisfactory way, would be considered as failures.

### Timescale

Time scale for this work is aligned with the time-frame for writing a Bachelor Thesis on Luleå University of Technology in the spring term of 2015. Progress reports will be sent to Capgemini on the dates of course seminaries where progress on the thesis is presented. These dates are as follows:

- 2015-01-22 PM seminary
- 2015-02-12 Seminary 1
- 2015-03-02 Seminary 2
- 2015-04-14 Seminary 3
- 2015-05-13 Paj Seminary
- 2015-06-** Final Presentation

The deadline for the whole project is June 2015. Upon evaluation of the results, a decision will be made about project's continuation.

## Key Constraints

### Organizational Constraints

- Because the Enterprise Architecture of Business Data Lake allows the policies to be formulated according to a specific user case, the policies formulation will not be the focus of this project. Instead, the focus will lie on EA itself and eventual architectural changes that could prove necessary to enable implementation of new GDPR requirements, including policies.
- Planning and managing the change of current enterprise governance and support models is a process that could really be started only after validation of produced artefacts and is therefore out of scope of this work. Thus, TOGAF ADM Phases will be performed up to Enterprise Architecture's design, with phase D as the final phase and Gap analysis as a final deliverable of design phases.

### Budget Information and Financial Constraints

This project is being done as Bachelor Thesis on Luleå University of Technology with a non-existent financial budget. The only things actually invested are work and time.

Expenses for interviews are also not a factor. Capgemini has generously provided and made available Christofer Holmgren Bagge as an expert advisor.

An interview planned with Neo4j graph database experts also comes at no expense. Upon completing an online course in Neo4j, a free pass for Neo4j conference in Stockholm 2015-03-17 was granted, with guaranteed one-on-one consultation time with one of the Neo4j experts.

### External and Business Constraints

No live instance of Business Data Lake is available for testing, so the project is going to be performed based on available documentation and interviews. For the same reason, validation of the results will be performed through consultation with Capgemini.

A possible constraint for implementation of "The Right to Be Forgotten" GDPR requirement is a phenomenon called immutability of Hadoop. An investigation will be performed into the matter to verify if this indeed is a constraint.

## Additional Information

### Current architecture/business/IT system descriptions or diagrams

At this point, descriptions and diagrams provided on the current architecture/business/IT system are:

- Capgemini, (2013), The Technology of the Business Data Lake, [http://www.capgemini.com/resource-file-access/resource/pdf/pivotal-business-data-lake-technical_brochure_web.pdf]
- Capgemini, (2013), The Technology Behind the Business Data Lake, [http://www.capgemini.com/resources/video/the-technology-behind-the-business-data-lake]
- Capgemini, (2015), The Business Data Lake: delivering the speed and accuracy to solve your big data problems, [http://www.capgemini.com/resource-file-access/resource/pdf/the_business_data_lake-delivering_the_speed_and_accuracy_to_solve_your_big_data_problems.pdf]
- Capgemini, Pivotal, (2014), Traditional BI vs. Business Data Lake – A comparison, [http://www.slideshare.net/capgemini/traditional-bi-vs-business-data-lake-a-comparison]

# Appendix E

## Stakeholder Map

| Stakeholder Map | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Stakeholder** | **Involvement** | **Class** | **Power** | **Level of Interest** | **Concerns / Requirements** | **Communication Plan** | **View** | **Artefacts/Deliverables** |
| Project sponsor | Client, the one who comissioned the development. This stakeholder group is interested in the high-level drivers, goals, and objectives of the organization, and how these are translated into an effective process and IT architecture to advance the business. | Keep Satisfied | High | High | Budgets, Demonstrable Benefit to the Enterprise, Business continuity, Business cost. | Indirect, through Project Management Group. | General views of all three architecture layers (Business, Information systems , and Technology), Data custodian's views of all three architecture layers, Data subject's views of all three architecture layers. | All produced artefacts and deliverables: Architecture Principles, Request for Architecture Work, Stakeholder Map, Business scenarios, Use cases, Requirements list, Assessment of Readiness for Business Transformation, Target Value propositions and KPIs, Business Transformation Risks' table, Baseline and Target Architecture Vision, Baseline Business architecture description, Target Business Architecture description, Baseline Information System's architectures description, Target Information System's architectures description, Baseline Technology architecture description, Target Technology Architecture description, Gap analysis, Solution design for Personal metadata repository, Prototype of Personal metadata repository. |
| Project Management Group | A business unit inside the project sponsor that is assigned with managing this project. This stakeholder group is interested in prioritizing, funding, and aligning change activity. An understanding of project content and technical dependencies improves portfolio management and decision-making. | Keep Satisfied | Medium | High | Success of the project, usability of results. | Planned progress reports on set dates. Detailed information is available in Request for Architecture Work document. Christofer Holmgren Bagge, the project's supervisor, is available as a contact person if the need arises for unplanned communication/consultation. | General views of all three architecture layers (Business, Information systems , and Technology), Data custodian's views of all three architecture layers, Data subject's views of all three architecture layers. | Same as for Project sponsor. |
| Users | This stakeholder group is interested in usability of Business Data Lake towards reaching their own high-level drivers, goals, and objectives of the organization, and how these are translated into an effective process and IT architecture to advance the business. This stakeholder group is familiar with the current system and will use the future system. Their own enerprise is a customer of the Project Sponsor. | Keep Satisfied | Medium | High | Demonstrable benefits to their own Enterprise, Business continuity, Business cost. The main concern is achieving compliance without the loss of system's usability towards achieving this stakeholder group's Business Goals. | Responsibility of Project sponsor or Project Management Group. No direct contact with users is enabled under the scope of this project. | General views of all three architecture layers (Business, Information systems , and Technology). | To be decided by the project sponsor after the completion of the project. |
| Competition | This stakeholder group is interested in solving the same compliance issue with GDPR as Business Data Lake product has. Therefore it is reasonable that they would be interested in the results of this work. | Responsibility of Project sponsor or Project Management Group | Low | Medium | Demonstrable Benefit to their own Enterprise, Business continuity, Business cost. The fact that a large part of the technology behind Business Data Lake is open source, makes all made improvements potentially available to competition as well. | Responsibility of Project sponsor or Project Management Group, presumably dictated by cooperation agreements and shared technology. | No Architecture Views are made specifically for this stakeholder group. | Responsibility of Project sponsor or Project Management Group. Presumably, the artefacts would need to be shared if they happen to fall under the scope of open source licences. |
| Data Protection Authorities | Government agencies. This stakeholder group is interested that the GDPR is enforced in enterprises and is therefore highly interested in if Business Data Lake Model is compliant with this regulation. | Keep Satisfied | High | High | Protection of personal data of EU citizens. Enforcement of GDPR on strategic level. | No communication plan at this point. Future communication plan is expected to be specified by DPAs themselves, upon their constitution. | To be specified after Data Protection Authorities are formed. Presumably all manufactured views on all layers. | Responsibility of Project sponsor or Project Management Group. The exact list of artefacts and deliverables requested is expected to be specified by DPAs themselves, upon their constitution. |
| Data custodians | Persons responsible for data management and auditing, more precisely Data Protection Officers, according to GDPR | Keep Informed | Medium | High | Making sure that personal data of EU citizens is managed in compliance with standing legislation. Enforcement of GDPR on tactical level. | General communication plan to this stakeholder group will be established by the Project Sponsor or the Project Management Group. Individual comunication plan with an expert covering this subject area is specified in the Interview Schedule document. | Data custodian's views on all three layers of architecture. | Baseline Business architecture description, Target Business Architecture description, Baseline Information System's architectures description, Target Information System's architectures description, Baseline Technology architecture description, Target Technology Architecture description, Gap analysis, Solution design for Personal metadata repository, Prototype of Personal metadata repository. |
| Software engineers and other technology experts | Technology experts involved in Business Data Lake development lifecycle and/or in development of related technologies. | Keep Informed | Medium | Medium | Incorporating changes caused by GDPR implementation into existing and future technologies. | General communication plan to this stakeholder group will be established by the Project Sponsor or the Project Management Group. Individual comunication plan with technology experts that are relevant for the accomplishment of this project is specified in the Interview Schedule document. | No Architecture Views are made specifically for this stakeholder group. Various architecture views may be of relevance based on the impact of the designed changes to supporting technologies. | Artefacts and deliverables of importance to the supporting technology's area may be available to technology experts that are involved in Business Data Lake development lifecycle and/or in development of related technologies. |
| Data subjects | Persons that the processed personal data refers to. | Keep Informed | Low | High | Making sure that personal data is managed in compliance with standing legislation. Right to be erased. Right to get a copy of the data that refers to them in an usable format. | No communication plan at this point. If the project ends succesfully, informing data subjects of the new product functionality would be consideret as a part of product marketing. | Data subject's view of target Business architecture. | Target architecture description - Data subject's view of target Business architecture. |

Table E1, Stakeholder map

# Appendix F

## Business Goals

**Improving Profitability** is a common business goal that is achieved by increasing income, lowering of expenses or both. The implementation of GDPR is expected to incur the cost of performing a system change, as well as the cost of additional data governance, but the potential cost of non-compliance is that the system becomes unusable for processing of personal data. Given the broad definition of personal data according to GDPR and proven downsides of data-set anonymization efforts, the usefulness of the current system to its current users would greatly diminish, resulting in probable decrease of sales and reduced income.

**Increase Market** Share is a business goal that is closely linked with profitability, which is usually proportionate to a market share. Changes of legislation that regulates the management of personal data can be seen as a market differentiator where those market-players that succeed in achieving the compliance of their products with the coming legislation are expected to maintain or increase their share of the market, while those whose product fail to satisfy the new compliance requirements are expected to lose their market share.

**Improving Company Image** is a business goal that relates to how the company is viewed in public, by its customers, potential customers, government and competition. Compliance with GDPR would further this goal by creating a public image of a company that cares about privacy of personal data. Furthermore, the image of a company that cares about its customers, by timely updating its products in response to changed requirements, would be stated.

**Improving Business Process Performance** is a business goal that, in relation to the issue of compliance with GDPR, means that the performance of business processes shall not be degraded by the system changes that would be necessary in order for compliance to be achieved.

**Improving Effectiveness of IT Organization** is a business goal that, in the scope of compliance with GDPR, means that the effectiveness of IT Organization shall not be degraded by the system changes that would be necessary in order for compliance to be achieved.

**Improving User Productivity** is a business goal that, in the scope of compliance with GDPR, means that user productivity would be improved or remains on the same level, following the system changes that are necessary in order for compliance to be achieved. The achievement of this goal could prove to be difficult because a greater share of data will now be classified as personal data, creating an overhead in data processing.

**Improving Portability and Scalability** is a business goal that, in the scope of compliance with GDPR, means that the inherent scalability of a Business Data Lake would not be decreased by the system changes that would be necessary in order for compliance to be achieved. Portability will be improved by satisfying GDPR requirement of providing a copy of personal data in an interchangeable format upon user’s request.

**Improving Interoperability** is a business goal that, in the scope of compliance with GDPR, means that the data exchange requires semantic interoperability. This is an important goal that must be achieved in order for personal data to be correctly classified as such. In practice, it means that metadata would have to be systematically collected and brought under the same standard.

**Decreasing Vendor Dependency** is a business goal that means that dependency on other vendors shall be decreased. In practice, this means that in-house and open-source solutions have the priority over closed-source solutions belonging to other vendors.

**Improving Security** is a business goal that, in the scope of compliance with GDPR, means the level of system and data security would be brought and/or maintained on a level required for protection of personal data.

## Business Drivers

**Improving Compliance** is the most immediate business driver for the system change that is the subject of this work. The need for improving compliance is created by the introduction of new legislation – the GDPR (General Data Protection Regulation). GDPR presents new, challenging requirements in the area of compliance and data governance that must be met in order for Business Data Lake architecture model to remain a competitive solution for managing Big Data. Improving compliance can even be seen as a business-critical issue because of the proposed penalties for non-compliance that go up to 5% of global turnover for the whole organization, limited to 100M Euro.

**Making Business Data Lake a leading solution to managing and processing Big Data** is the main business driver. The opportunity for accomplishing this is seen in using compliance issue with GDPR as a differentiator that would, after successfully achieving the compliance, create a competitive advantage against competing solutions that did not manage to overcome this issue.

## Organizational Constraints

**The policies formulation will not be the focus of this project**. The reason for this constraint is that Business Data Lake architecture model allows the policies to be formulated according to a specific user's requirements. This means that policies could differ significantly depending on the intended usage. Instead, the focus will lie on EA itself and eventual architectural changes that could prove necessary to enable implementation of new GDPR requirements, including policies.

**Planning and managing the change of current enterprise governance and support models is out of scope of this project**. The reason for this constraint is that such a process could really be started only after validation of produced artefacts and deliverables. Therefore, TOGAF ADM Phases will be performed up to completion of Enterprise Architecture's design – phase D.

## Financial Constraints

**Financial budget for this project is non-existent.** This project is being done as Bachelor Thesis on Luleå University of Technology. The only things actually invested are work and time.

**Expenses for interviews are also non-existent**. Capgemini has generously provided and made available Christofer Holmgren Bagge, Managing Consultant and Senior Business Analyst with focus on Big Data and Big Data Analytics at Capgemini Sweden as an expert advisor. An interview with Neo4j graph database expert also came at no expense. Upon completing an online course in Neo4j, a free pass for Neo4j conference in Stockholm 2015-03-17 was granted, on which a one-hour consultation time with Jim Webber, Ph.D., Chief Scientist at Neo4j, was procured thanks to generosity of Neo4j.

## External and Business Constraints

**No live instance of Business Data Lake is available for prototyping.** For this reason the project is going to be performed based on available documentation and interviews. The validation of the results will be performed through consultation with project's sponsor - Capgemini.

**Implementation of "The Right to Be Erased" GDPR requirement may be dependent on the inherited constraint of Hadoop.** This constraint is a phenomenon called immutability of Hadoop that hinders the insurance of physical deletion of data. An investigation will be performed into the matter to verify if this indeed is a constraint that cannot be overcome.

# Appendix G

## Baseline Use Cases

### <u>Business user</u>

*UC1:*

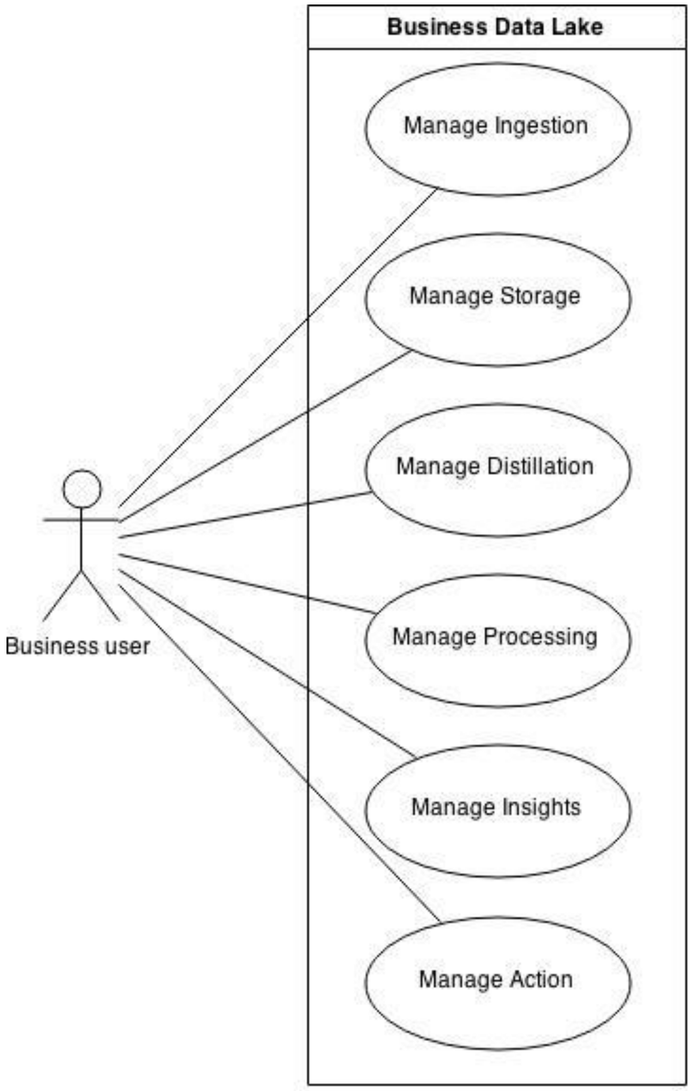


Figure G1, USE CASE 1

| | |
|---|---|
| Name: | UC1 |
| Summary: | This is a high-level abstraction of business user’s use case. The Business user manages all business operations on data through Unified Operations Environment. |
| Scope: | Business Data Lake |
| Actors: | Business user |
| Preconditions: | Business user has access rights to the system. |
| Aftermath: | Business user has finished intended operations on data and has logged out. |
| Success Scenario: | 1. Business user logs in to Unified Operations Environment.<br>2. Business user chooses the process where his intended business operation is to take place and a sandbox environment is created. |

| | |
|---|---|
| | 3. Business user chooses the data that he has the access rights to and the copy of this data is transferred to the sandbox environment.<br>4. Business user performs the operation.<br>5. Business user performs further operations that would ultimately lead to business action.<br>6. Business action is performed.<br>7. The user logs out.<br>8. The newly generated data and metadata is returned to the data lake. |
| Extensions: | From step 1:<br>1. Login credentials are refused, a message is displayed.<br>2. Business user is returned to step 1.<br>From step 4:<br>1. Business user logs out.<br>2. The newly generated data and metadata is returned to the data lake. |

Table G1, USE CASE 1

*UC2:*

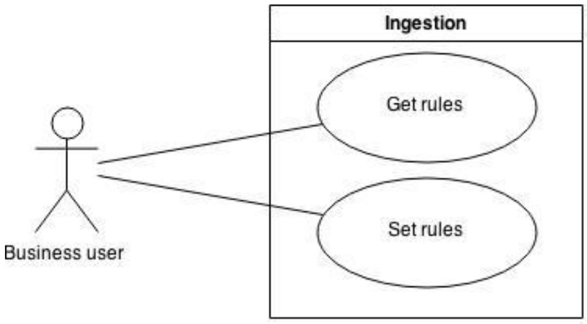


Figure G2, USE CASE 2

| | |
|---|---|
| Name: | UC2 |
| Summary: | Business user reviews and chooses rules for ingestion of data. |
| Scope: | Ingestion |
| Actors: | Business user |
| Preconditions: | Business user has logged in on Unified Operations Environment and chooses ingestion process as a starting point for his intended business operation. |
| Aftermath: | Business user has finished intended operations and has logged out. |
| Success Scenario: | 1. Business user reviews ingestion rules for data of interest.<br>2. Business user sets new ingestion rules for data of interest.<br>3. Business user logs out. |
| Extensions: | From step 2:<br>2. The setting of new ingestion rules is refused on account of insufficient access rights, a message is displayed.<br>3. User is returned to step 1. |

Table G2, USE CASE 2

*UC3:*

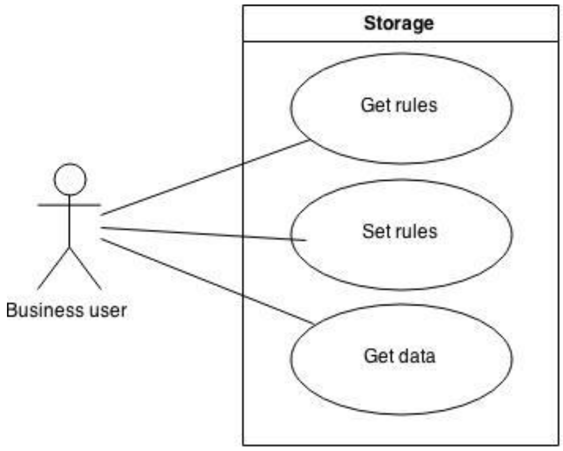


Figure G3, USE CASE 3

| | |
|---|---|
| Name: | UC3 |
| Summary: | Business user reviews and chooses rules for storage of data. Business user chooses the data for performing business operations on. |
| Scope: | Storage |
| Actors: | Business user |
| Preconditions: | Business user has logged in on Unified Operations Environment and chooses the storage process as a starting point for his intended business operation. |
| Aftermath: | The chosen data is transferred to the sandbox environment for structuring and/or analysis. |
| Success Scenario: | 1. Business user reviews storage rules for data of interest.<br>2. Business user sets new storage rules for data of interest.<br>3. Business user chooses the data that he has the access rights to and the copy of this data is transferred to the sandbox environment for structuring and/or analysis. |
| Extensions: | From step 2:<br>2. The setting of new storage rules is refused on account of insufficient access rights, a message is displayed.<br>3. User is returned to step 1.<br>From step 2:<br>2. Business user chooses the data that he has the access rights to and the copy of this data is transferred to the sandbox environment for structuring and/or analysis.<br>From step 3:<br>3. Transfer of data is refused on account of insufficient access rights, a message is displayed.<br>4. User is returned to step 3. |

Table G3, USE CASE 3

*UC4:*

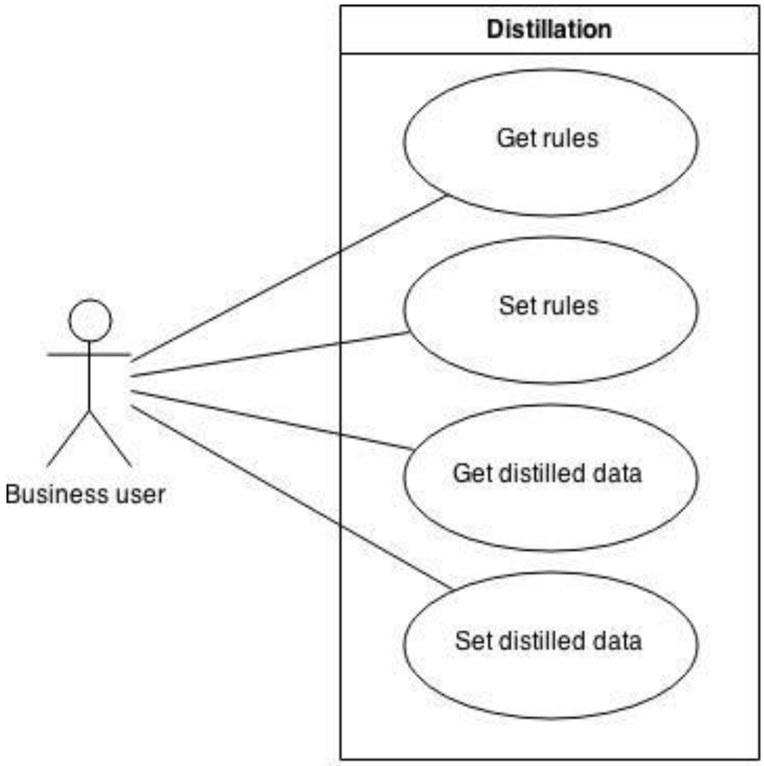


Figure G4, USE CASE 4

|  |  |
|---|---|
| Name: | UC4 |
| Summary: | Business user reviews and chooses rules for distillation of data, reviews and modifies distilled data when needed. Business user starts business operations on distilled data. |
| Scope: | Distillation |
| Actors: | Business user |
| Preconditions: | Business user has logged in on Unified Operations Environment and chooses the distillation process as the environment for his intended business operation or for continuance of an already started business operation. |
| Aftermath: | Business user has finished intended operations and has logged out. Alternatively, the user has proceeded with further processing and/or analysis of data. |
| Success Scenario: | 1. Business user reviews distillation rules for data of interest.<br>2. Business user sets new distillation rules for data of interest.<br>3. Business user reviews the distilled data.<br>4. The user logs out. |
| Extensions: | From step 2:<br>2. The setting of new distillation rules is refused on account of insufficient access rights, a message is displayed.<br>3. User is returned to step 1.<br>From step 4:<br>4. Business user proceeds with further processing and/or analysis of data.<br>From step 4:<br>4. Business user modifies distilled data.<br>5. The user logs out. |

| | |
|---|---|
| | From step 4:<br>4. Business user modifies distilled data.<br>5. Business user proceeds with further processing and/or analysis of data. |

Table G4, USE CASE 4

*UC5:*

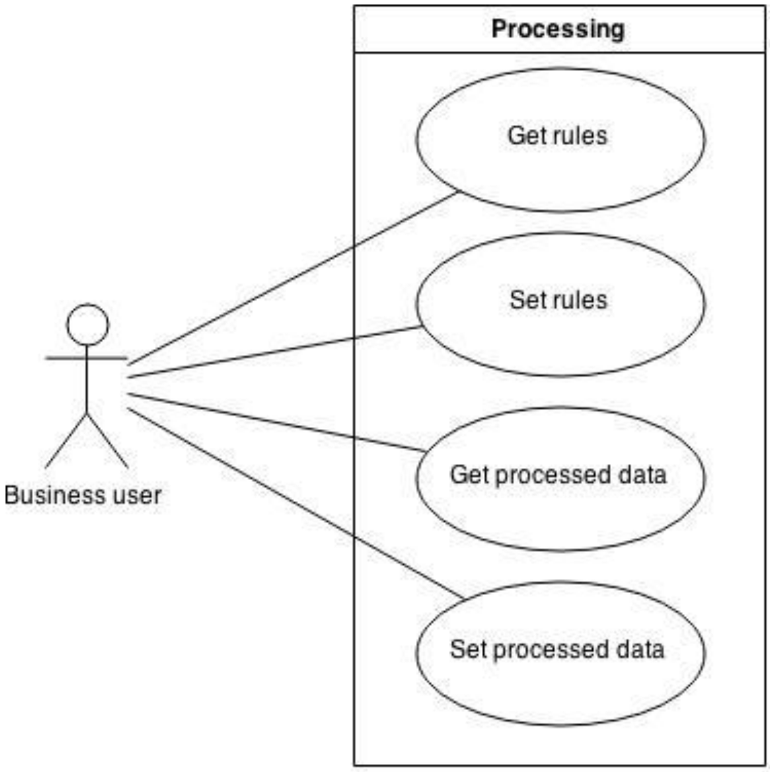


Figure G5, USE CASE 5

| | |
|---|---|
| Name: | UC5 |
| Summary: | Business user reviews and chooses rules for processing of data, reviews and modifies processed data when needed. Business user starts business operations on processed data. |
| Scope: | Processing |
| Actors: | Business user |
| Preconditions: | Business user has logged in on Unified Operations Environment and chooses processing as the environment for his intended business operation or for continuance of an already started business operation. |
| Aftermath: | Business user has finished intended operations and has logged out. Alternatively, the user has proceeded with further processing and/or analysis of data. |
| Success Scenario: | 1. Business user reviews processing rules for data of interest.<br>2. Business user sets new processing rules for data of interest.<br>3. Business user reviews the processed data.<br>4. The user logs out. |
| Extensions: | From step 2:<br>2. The setting of new processing rules is refused on account of insufficient access rights, a message is displayed.<br>3. User is returned to step 1. |

| | |
|---|---|
| | From step 4:<br>4. Business user proceeds with further processing and/or analysis of data.<br>From step 4:<br>5. Business user modifies the processed data.<br>6. The user logs out.<br>From step 4:<br>4. Business user modifies the processed data.<br>5. Business user proceeds with further processing and/or analysis of data. |

Table G5, USE CASE 5

*UC6:*

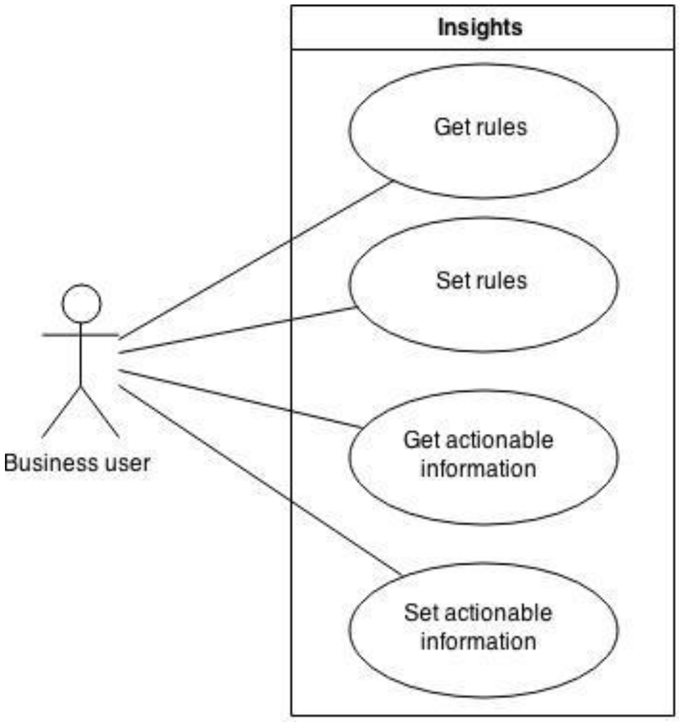


Figure G6, USE CASE 6

| | |
|---|---|
| Name: | UC6 |
| Summary: | Business user reviews and chooses rules for creation of insights, reviews and modifies insights when needed. Business user starts business operations on insights. |
| Scope: | Insights |
| Actors: | Business user |
| Preconditions: | Business user has logged in on Unified Operations Environment and chooses insights as the environment for his intended business operation or for continuance of an already started business operation. |
| Aftermath: | Business user has finished intended operations and has logged out. Alternatively, the user has proceeded with processing and/or analysis of insights. |
| Success Scenario: | 1. Business user reviews insight rules for data of interest.<br>2. Business user sets new insight rules for data of interest.<br>3. Business user reviews the insights.<br>4. The user logs out. |

| | |
|---|---|
| Extensions: | From step 2:<br>2. The setting of new insight rules is refused on account of insufficient access rights, a message is displayed.<br>3. User is returned to step 1.<br>From step 4:<br>4. Business user proceeds with further processing and/or analysis of data.<br>From step 4:<br>4. Business user modifies the insights.<br>5. The user logs out.<br>From step 4:<br>4. Business user modifies the insights.<br>5. Business user proceeds with further processing and/or analysis of insights. |

Table G6, USE CASE 6

*UC7:*

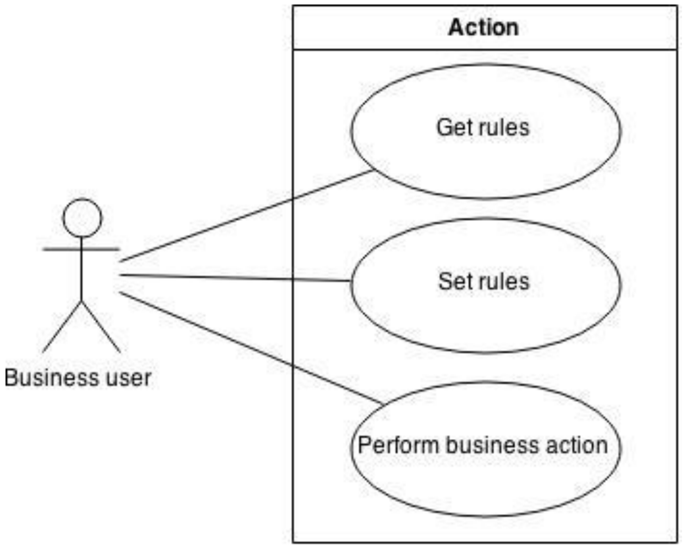


Figure G7, USE CASE 7

| | |
|---|---|
| Name: | UC7 |
| Summary: | Business user reviews and chooses rules for creation of business actions, performs business actions. |
| Scope: | Action |
| Actors: | Business user |
| Preconditions: | Business user has logged in on Unified Operations Environment and chooses action as the environment for his intended business operation or for continuance of an already started business operation. |
| Aftermath: | Business user has finished intended operations and has logged out. |
| Success Scenario: | 1. Business user reviews action rules for data of interest.<br>2. Business user sets new action rules for data of interest.<br>3. Business user performs business action.<br>4. The user logs out. |
| Extensions: | From step 2:<br>2. Business user proceeds with further processing and/or analysis of data. |

| | |
|---|---|
| | From step 3:<br>3. The user logs out. |

Table G7, USE CASE 7

## Data custodian

*UC8:*

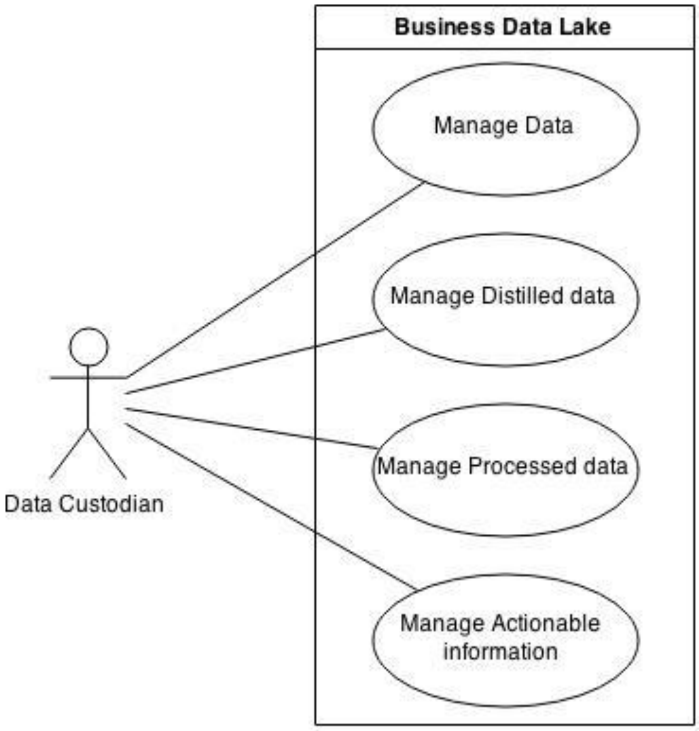


Figure G8, USE CASE 8

| | |
|---|---|
| Name: | UC8 |
| Summary: | This is a high-level abstraction of data custodian's use case. The Data custodian manages all the data stored across the system through the Unified Data Management Environment. |
| Scope: | Business Data Lake |
| Actors: | Data custodian |
| Preconditions: | Data custodian has the appropriate access rights to the Unified Data Management Environment. |
| Aftermath: | Data custodian has finished intended management of data and has logged out. |
| Success Scenario: | 1. Data custodian logs in to Unified Data Management Environment.<br>2. Data custodian chooses the data, distilled data, processed data or actionable information that needs to be managed.<br>3. Data custodian performs the management.<br>4. New access rules and new metadata are created.<br>5. Data custodian logs out. |
| Extensions: | From step 1:<br>1. Login credentials are refused, a message is displayed.<br>2. Business user is returned to step 1. |

Table G8, USE CASE 8

*UC9:*

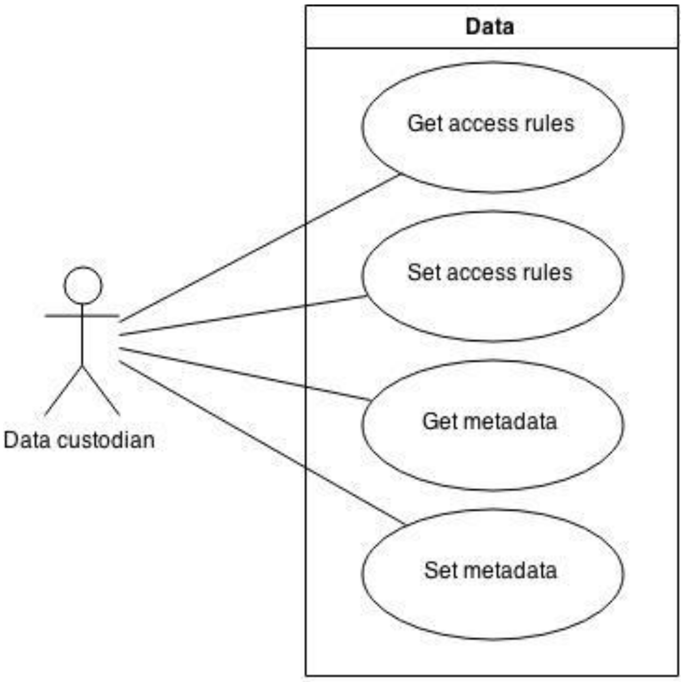


Figure G9, USE CASE 9

| | |
|---|---|
| Name: | UC9 |
| Summary: | Data custodian manages stored data. |
| Scope: | Data |
| Actors: | Data custodian |
| Preconditions: | Data custodian has logged in on Unified Data Management Environment and chooses Data as the target for intended management. |
| Aftermath: | Data custodian has finished intended management of data and has logged out. |
| Success Scenario: | 1. Data custodian chooses the data as the target of intended management.<br>2. Data custodian reviews access rule for the data.<br>3. Data custodian updates access rules for the data.<br>4. Data custodian reviews metadata.<br>5. Data custodian updates metadata.<br>6. Data custodian logs out. |
| Extensions: | From step 2:<br>2. Order of steps (2,3) and (3,4) switched.<br>From step 2:<br>2. Data custodian reviews metadata.<br>3. Data custodian updates metadata.<br>4. Data custodian logs out.<br>From step 3:<br>3. Data custodian logs out. |

Table G9, USE CASE 9

*UC10:*

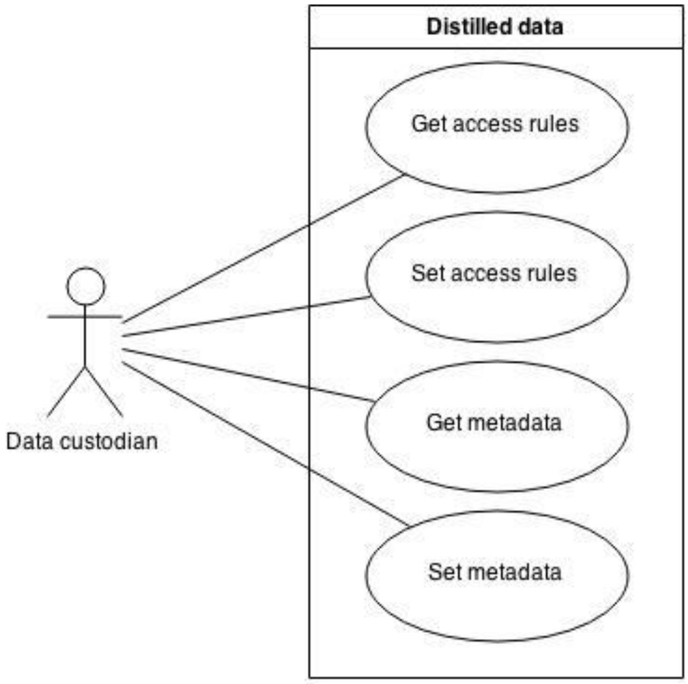


Figure G10, USE CASE 10

| | |
|---|---|
| Name: | UC10 |
| Summary: | Data custodian manages distilled data. |
| Scope: | Distilled data |
| Actors: | Data custodian |
| Preconditions: | Data custodian has logged in on Unified Data Management Environment and chooses Distilled data as the target for intended management. |
| Aftermath: | Data custodian has finished intended management of data and has logged out. |
| Success Scenario: | 1. Data custodian chooses the Distilled data as the target of intended management.<br>2. Data custodian reviews access rule for the distilled data.<br>3. Data custodian updates access rules for the distilled data.<br>4. Data custodian reviews metadata.<br>5. Data custodian updates metadata.<br>6. Data custodian logs out. |
| Extensions: | From step 2:<br>2. Order of steps (2,3) and (3,4) switched.<br>From step 2:<br>2. Data custodian reviews metadata.<br>3. Data custodian updates metadata.<br>4. Data custodian logs out.<br>From step 3:<br>3. Data custodian logs out. |

Table G10, USE CASE 10

*UC11:*

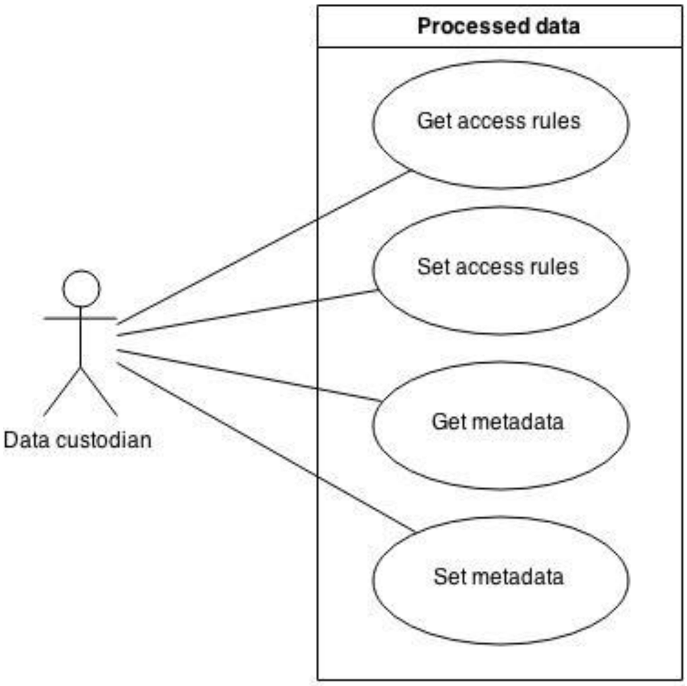


Figure G11, USE CASE 11

| | |
|---|---|
| Name: | UC11 |
| Summary: | Data custodian manages Processed data. |
| Scope: | Processed data |
| Actors: | Data custodian |
| Preconditions: | Data custodian has logged in on Unified Data Management Environment and chooses Processed data as the target for intended management. |
| Aftermath: | Data custodian has finished intended management of data and has logged out. |
| Success Scenario: | 1. Data custodian chooses the processed data as the target of intended management.<br>2. Data custodian reviews access rule for the processed data.<br>3. Data custodian updates access rules for the processed data.<br>4. Data custodian reviews metadata.<br>5. Data custodian updates metadata.<br>6. Data custodian logs out. |
| Extensions: | From step 2:<br>2. Order of steps (2,3) and (3,4) switched.<br>From step 2:<br>2. Data custodian reviews metadata.<br>3. Data custodian updates metadata.<br>4. Data custodian logs out.<br>From step 3:<br>3. Data custodian logs out. |

Table G11, USE CASE 11

*UC12:*

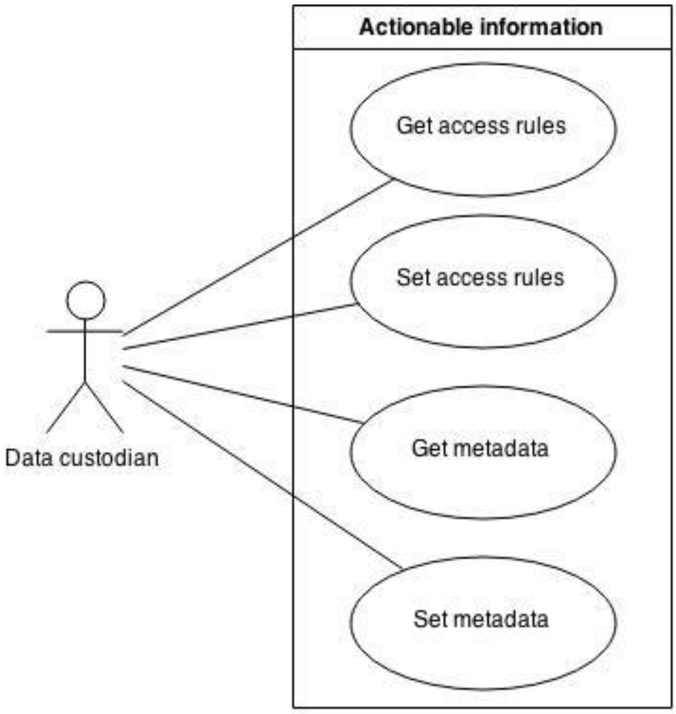


Figure G12, USE CASE 12

| | |
|---|---|
| Name: | UC12 |
| Summary: | Data custodian manages Actionable information. |
| Scope: | Actionable information |
| Actors: | Data custodian |
| Preconditions: | Data custodian has logged in on Unified Data Management Environment and chooses Actionable information as the target for intended management. |
| Aftermath: | Data custodian has finished intended management of data and has logged out. |
| Success Scenario: | 7. Data custodian chooses the actionable information as the target of intended management.<br>8. Data custodian reviews access rule for the actionable information.<br>9. Data custodian updates access rules for the actionable information.<br>10. Data custodian reviews metadata.<br>11. Data custodian updates metadata.<br>12. Data custodian logs out. |
| Extensions: | From step 2:<br>2. Order of steps (2,3) and (3,4) switched.<br>From step 2:<br>2. Data custodian reviews metadata.<br>3. Data custodian updates metadata.<br>4. Data custodian logs out.<br>From step 3:<br>3. Data custodian logs out. |

Table G12, USE CASE 12

## Data subject

*UC13:*

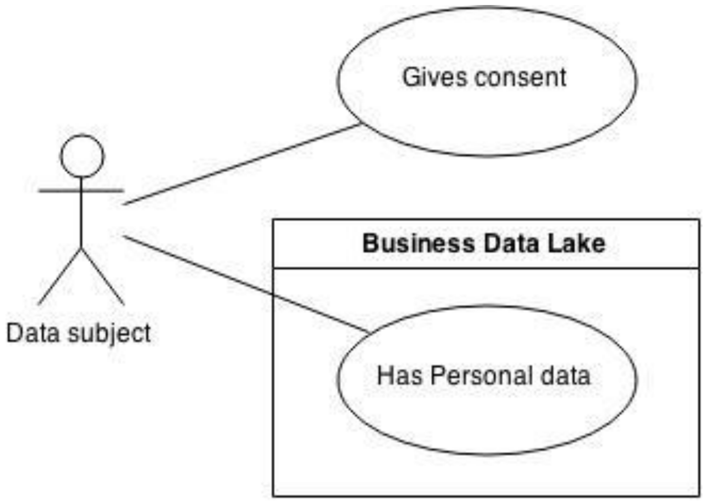


Figure G13, USE CASE 13

| | |
|---|---|
| Name: | UC3 |
| Summary: | Data subject gives consent for his personal data to be collected, shared, processed, etc. This data is ingested into the Business Data Lake. |
| Scope: | Business Data Lake |
| Actors: | Data subject |
| Preconditions: | During the extraction of data from data sources, Data subject has given his consent for his data to be gathered, shared and processed in practically any way possible. A note must be made that this is done outside the Business Data Lake system, but it is a safe assumption that no personal data is ingested without the consent of the data subject. |
| Aftermath: | Data subject's personal data is stored inside the Business Data Lake. Information about the given consent is presumably stored inside the metadata. |
| Success Scenario: | 1. Data subject gives consent for his personal data to be gathered, shared, processed, etc.<br>2. Data subject's personal data is ingested and stored inside the Business Data Lake. |
| Extensions: | From step 1:<br>1. Data subject does not give consent for his personal data to be gathered, shared, processed, etc.<br>2. Data subject's personal data is not ingested and stored inside the Business Data Lake. |

Table G13, USE CASE 13

## Target Use Cases

### Business user

There are no new or changed use cases involving Business user in Target Architecture Vision.

### Data custodian

*UC14:*

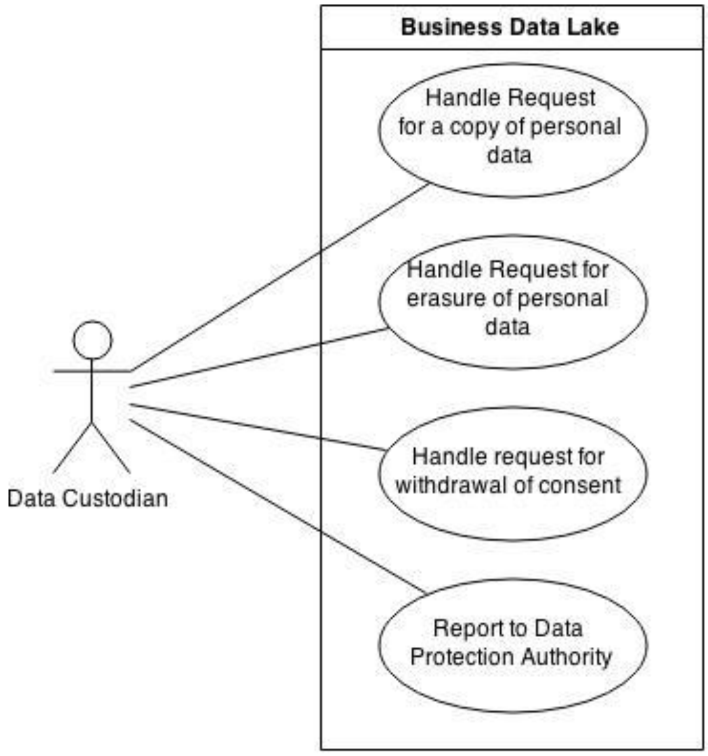

Business Data Lake
Handle Request for a copy of personal data
Handle Request for erasure of personal data
Handle request for withdrawal of consent
Report to Data Protection Authority
Data Custodian

Figure G14, USE CASE 14

| | |
|---|---|
| Name: | UC14 |
| Summary: | This is a high-level abstraction of data custodian's use case. The Data custodian manages all personal data stored across the system through the Unified Data Management Environment. |
| Scope: | Business Data Lake |
| Actors: | Data custodian |
| Preconditions: | Data custodian has the appropriate access rights to the Unified Data Management Environment. |
| Aftermath: | Data custodian has finished intended management of personal data and has logged out. |
| Success Scenario: | 1. Data custodian logs in to Unified Data Management Environment.<br>2. Data custodian chooses the personal data that needs to be managed amongst data, distilled data, processed data or actionable information.<br>3. Data custodian performs the management.<br>4. Data custodian logs out. |
| Extensions: | From step 1:<br>1. Login credentials are refused, a message is displayed. |

| | |
|---|---|
| | 2. Data custodian is returned to step 1. |

Table G14, USE CASE 14

*UC15:*

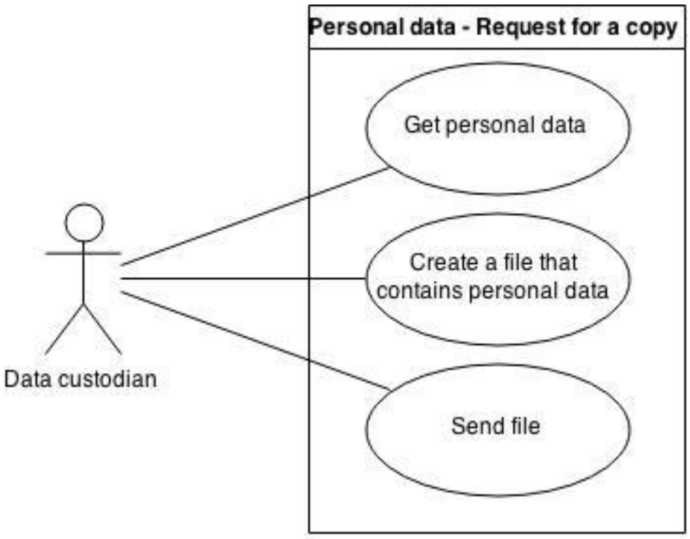

Personal data - Request for a copy
Get personal data
Create a file that contains personal data
Send file
Data custodian

Figure G15, USE CASE 15

| | |
|---|---|
| Name: | UC15 |
| Summary: | Data custodian handles the Request for a copy of personal data that is made by Data subject. |
| Scope: | Personal data |
| Actors: | Data custodian |
| Preconditions: | Data custodian has received the Request for a copy of personal data that was sent by a Data subject. |
| Aftermath: | Data custodian has sent the copy of requested personal data to the Data subject who sent the request. |
| Success Scenario: | 1. Data custodian logs in to Unified Data Management Environment.<br>2. Data custodian selects the personal data that relates to the Data subject.<br>3. Data custodian copies the selected data into a file.<br>4. Data custodian sends the file to the Data subject.<br>5. Data custodian logs out. |
| Extensions: | From step 3:<br>3. No personal data relating to the Data subject is found.<br>4. Data custodian notifies the Data subject about failing to find any personal data relating to the Data subject.<br>5. Data custodian logs out.<br>From step 1:<br>1. Login credentials are refused, a message is displayed.<br>2. Data custodian is returned to step 1. |

Table G15, USE CASE 15

*UC16:*

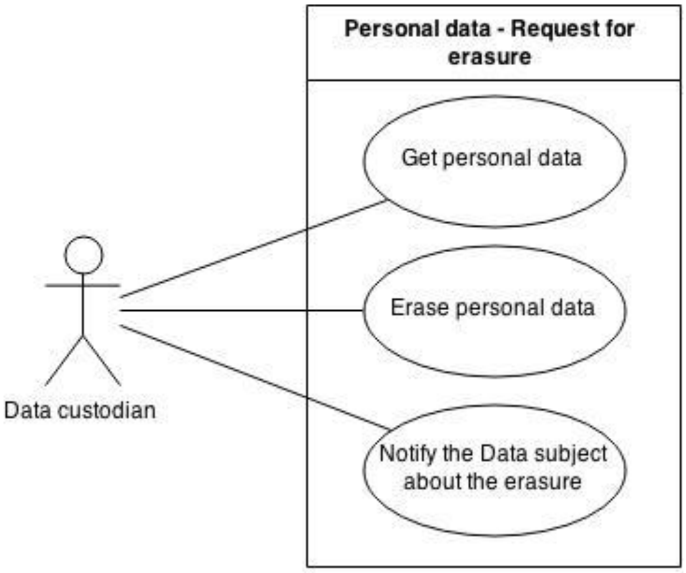

Personal data - Request for erasure
Get personal data
Erase personal data
Notify the Data subject about the erasure
Data custodian

Figure G16, USE CASE 16

| | |
|---|---|
| Name: | UC16 |
| Summary: | Data custodian handles the Request for erasure of personal data that is made by Data subject. |
| Scope: | Personal data |
| Actors: | Data custodian |
| Preconditions: | Data custodian has received the Request for erasure of personal data that was sent by a Data subject. |
| Aftermath: | Data custodian has erased all personal data related to the Data subject who sent the request. |
| Success Scenario: | 1. Data custodian logs in to Unified Data Management Environment.<br>2. Data custodian selects the personal data that relates to the Data subject.<br>3. Data custodian erases the selected data.<br>4. Data custodian notifies the Data subject about the erasure.<br>5. Data custodian logs out. |
| Extensions: | From step 3:<br>3. No personal data relating to the Data subject is found.<br>4. Data custodian notifies the Data subject about failing to find any personal data relating to the Data subject.<br>5. Data custodian logs out.<br>From step 1:<br>1. Login credentials are refused, a message is displayed.<br>2. Data custodian is returned to step 1. |

Table G16, USE CASE 16

*UC17:*

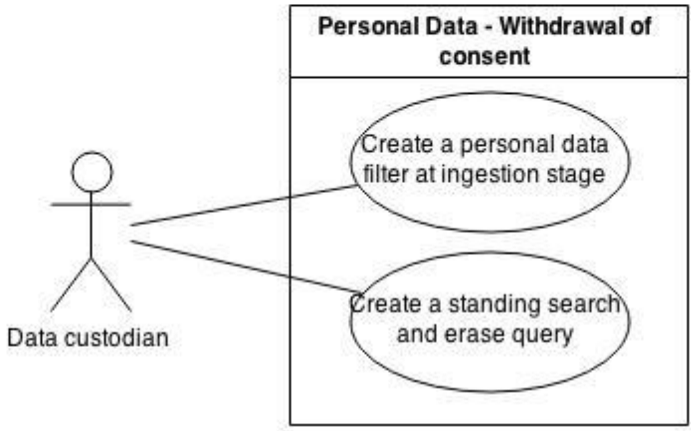

Personal Data - Withdrawal of consent
Create a personal data filter at ingestion stage
Create a standing search and erase query
Data custodian

Figure G17, USE CASE 17

| | |
|---|---|
| Name: | UC17 |
| Summary: | Data custodian handles the request for withdrawal of consent. |
| Scope: | Personal data |
| Actors: | Data custodian |
| Preconditions: | Data custodian has received the Withdrawal of consent for collecting and using Data Subject's personal data. |
| Aftermath: | Data custodian has set constraints for storing and using (double insurance) Data subject's personal data that is ingested after the arrival of the consent withdrawal request. |
| Success Scenario: | 1. Data custodian logs in to Unified Data Management Environment.<br>2. Data custodian creates a standing rule that filters out the personal data specified in the request at ingestion stage and prohibits its storage.<br>3. Data custodian creates a standing query that filters out the personal data specified in the request from storage stage and erases it.<br>4. Data custodian logs out. |
| Extensions: | From step 1:<br>1. Login credentials are refused, a message is displayed.<br>2. Data custodian is returned to step 1. |

Table G17, USE CASE 17

*UC18:*

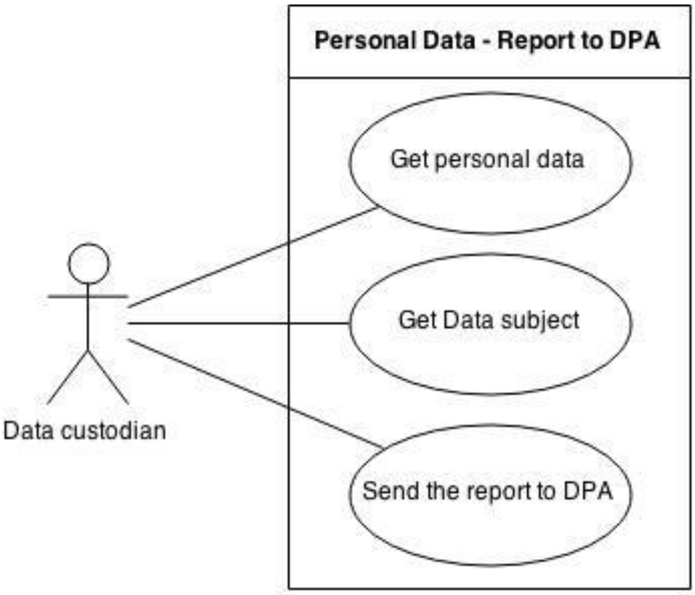


Figure G18, USE CASE 18

| | |
|---|---|
| Name: | UC18 |
| Summary: | Data custodian reports to Data Protection Authority. |
| Scope: | Personal data |
| Actors: | Data custodian |
| Preconditions: | A breach of security of personal data or/and breach of personal data management's compliance with GDPR has occurred and came to the knowledge of Data custodian. Data custodian has a responsibility to report this occurrence to Data Protection Authority. |
| Aftermath: | Data custodian has sent the report to DPA describing the occurrence, personal data and Data subjects that were affected. |
| Success Scenario: | 1. Data custodian logs in to Unified Data Management Environment.<br>2. Data custodian selects the personal data affected by the breach.<br>3. Data custodian selects the Data subjects affected by the breach.<br>4. Data custodian sends the report to the DPA.<br>5. Data custodian logs out. |
| Extensions: | From step 2:<br>2. No personal data was found to be affected by the breach.<br>3. Data custodian sends the report to the DPA.<br>4. Data custodian logs out.<br>From step 1:<br>1. Login credentials are refused, a message is displayed.<br>2. Data custodian is returned to step 1. |

Table G18, USE CASE 18

## Data subject

*UC19:*

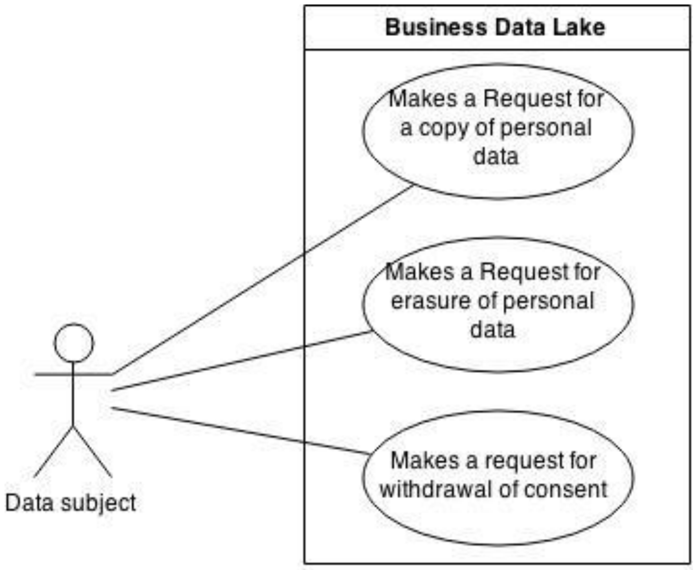

Business Data Lake
Makes a Request for a copy of personal data
Makes a Request for erasure of personal data
Makes a request for withdrawal of consent
Data subject

Figure G19, USE CASE 19

| | |
|---|---|
| Name: | UC19 |
| Summary: | This is a high-level abstraction of data subject’s use case. The Data subject exercises his newly acquired rights by the courtesy of GDPR. |
| Scope: | Personal data |
| Actors: | Data subject |
| Preconditions: | Data subject wishes to exercise his rights regarding his personal data. |
| Aftermath: | An organization is notified of Data subject’s request(s). |
| Success Scenario: | 1. Data subject makes a Request for a copy of personal data.<br>2. Data subject makes a Request for erasure of personal data.<br>3. Data subject makes a request for Withdrawal of consent |
| Extensions: | |

Table G19, USE CASE 19

*UC20:*

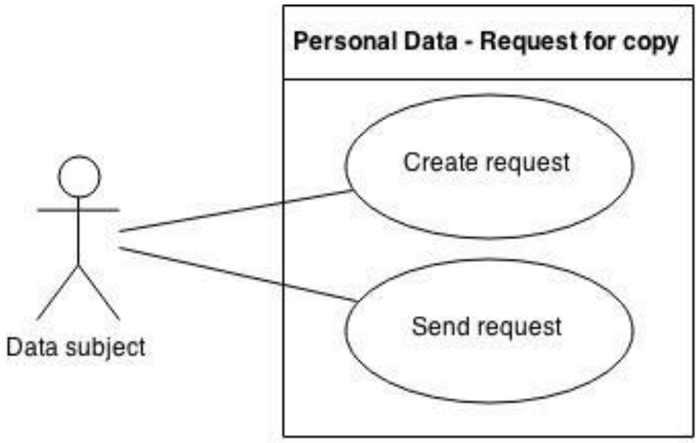


Figure G20, USE CASE 20

| | |
|---|---|
| Name: | UC20 |

| | |
|---|---|
| Summary: | Data subject makes a Request for a copy of personal data. |
| Scope: | Personal data |
| Actors: | Data subject |
| Preconditions: | Data subject wishes to get a copy of the data that an organization has on him. |
| Aftermath: | A Request for a copy of personal data is sent to Data custodian. |
| Success Scenario: | 4. Data subject creates the Request for a copy of personal data.<br>5. Data subject sends the Request for a copy of personal data to Data custodian. |
| Extensions: | |

Table G20, USE CASE 20

*UC21:*

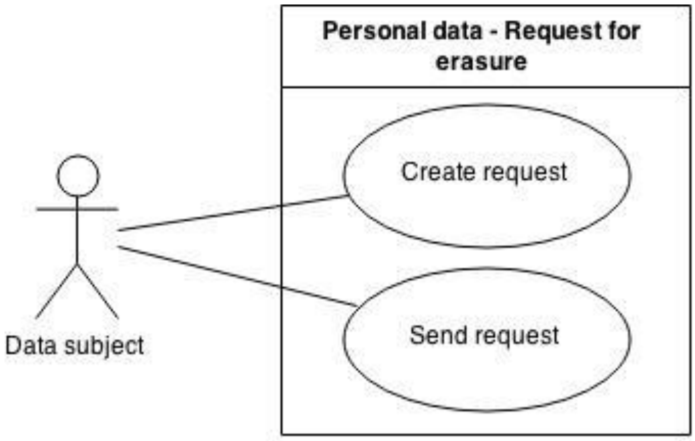


Figure G21, USE CASE 21

| | |
|---|---|
| Name: | UC21 |
| Summary: | Data subject makes a Request for erasure of personal data. |
| Scope: | Personal data |
| Actors: | Data subject |
| Preconditions: | Data subject wishes that an organization erases all the data it has on him. |
| Aftermath: | A Request for erasure of personal data is sent to Data custodian. |
| Success Scenario: | 1. Data subject creates the Request for erasure of personal data.<br>2. Data subject sends the Request for erasure of personal data to Data custodian. |
| Extensions: | |

Table G21, USE CASE 21

*UC22:*

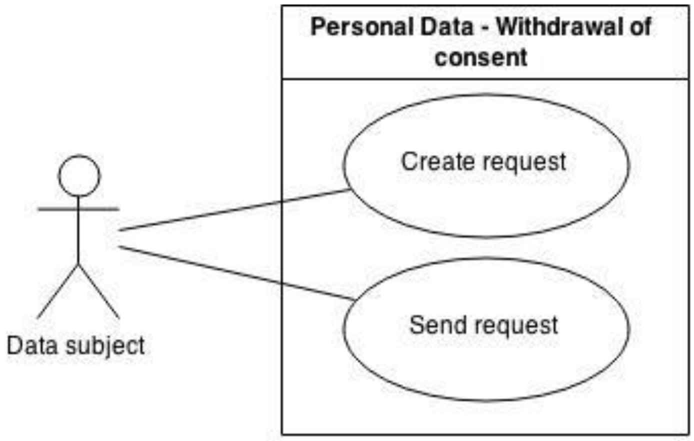


Figure G22, USE CASE 22

| | |
|---|---|
| Name: | UC22 |
| Summary: | Data subject makes a Withdrawal of consent request. |
| Scope: | Personal data |
| Actors: | Data subject |
| Preconditions: | Data subject wishes to withdraw his consent for collection and usage of his personal data by an organization. |
| Aftermath: | A Withdrawal of consent request is sent to Data custodian. |
| Success Scenario: | 1. Data subject creates a Withdrawal of consent request.<br>2. Data subject sends the Withdrawal of consent request to Data custodian. |
| Extensions: | |

Table G22, USE CASE 22

# Appendix H

## Baseline Architecture Vision's Functional Requirements

*BR01:*

| | |
|---|---|
| Requirement: | BR01 |
| Summary: | Business user has the capability to monitor, configure, and manage the whole platform by using a single unified operating environment. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H1, BASELINE REQUIREMENT 01

*BR02:*

| | |
|---|---|
| Requirement: | BR02 |
| Summary: | Business user's access to system's functionality and data is limited by access rights. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H2, BASELINE REQUIREMENT 02

*BR03:*

| | |
|---|---|
| Requirement: | BR03 |
| Summary: | Business user can review the rules of ingestion process. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H3, BASELINE REQUIREMENT 03

*BR04:*

| | |
|---|---|
| Requirement: | BR04 |
| Summary: | Business user can set the rules of ingestion process. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H4, BASELINE REQUIREMENT 04

*BR05:*

| | |
|---|---|
| Requirement: | BR05 |
| Summary: | Business user has read access to data. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H5, BASELINE REQUIREMENT 05

*BR06:*

| | |
|---|---|
| Requirement: | BR06 |
| Summary: | Business user can review the rules of distillation process. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H6, BASELINE REQUIREMENT 06**

*BR07:*

| | |
|---|---|
| Requirement: | BR07 |
| Summary: | Business user can set the rules of distillation process. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H7, BASELINE REQUIREMENT 07**

*BR08:*

| | |
|---|---|
| Requirement: | BR08 |
| Summary: | Business user has read access to distilled data. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H8, BASELINE REQUIREMENT 08**

*BR09:*

| | |
|---|---|
| Requirement: | BR09 |
| Summary: | Business user has write access to distilled data. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H9, BASELINE REQUIREMENT 09**

*BR10:*

| | |
|---|---|
| Requirement: | BR10 |
| Summary: | Business user can review the rules of processing. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H10, BASELINE REQUIREMENT 10**

*BR11:*

| | |
|---|---|
| Requirement: | BR11 |
| Summary: | Business user can set the rules of processing. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H11, BASELINE REQUIREMENT 11**

*BR12:*

| | |
|---|---|
| Requirement: | BR12 |
| Summary: | Business has read access to processed data. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H12, BASELINE REQUIREMENT 12

*BR13:*

| | |
|---|---|
| Requirement: | BR13 |
| Summary: | Business has write access to processed data. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H13, BASELINE REQUIREMENT 13

*BR14:*

| | |
|---|---|
| Requirement: | BR14 |
| Summary: | Business user can review the rules of insights process. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H14, BASELINE REQUIREMENT 14

*BR15:*

| | |
|---|---|
| Requirement: | BR15 |
| Summary: | Business user can set the rules of insights process. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H15, BASELINE REQUIREMENT 15

*BR16:*

| | |
|---|---|
| Requirement: | BR16 |
| Summary: | Business has read access to actionable information. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H16, BASELINE REQUIREMENT 16

*BR17:*

| | |
|---|---|
| Requirement: | BR17 |
| Summary: | Business has write access to actionable information. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H17, BASELINE REQUIREMENT 17

*BR18:*

| | |
|---|---|
| Requirement: | BR18 |
| Summary: | Business user can review the rules of action process. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H18, BASELINE REQUIREMENT 18**

*BR19:*

| | |
|---|---|
| Requirement: | BR19 |
| Summary: | Business user can set the rules of action process. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H19, BASELINE REQUIREMENT 19**

*BR20:*

| | |
|---|---|
| Requirement: | BR20 |
| Summary: | Data custodian has the capability to manage data through a single unified operating environment. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H20, BASELINE REQUIREMENT 20**

*BR21:*

| | |
|---|---|
| Requirement: | BR21 |
| Summary: | Data custodian's access to system's functionality, data and data management is limited by access rights. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H21, BASELINE REQUIREMENT 21**

*BR22:*

| | |
|---|---|
| Requirement: | BR22 |
| Summary: | Data custodian can review the access rules for data. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H22, BASELINE REQUIREMENT 22**

*BR23:*

| | |
|---|---|
| Requirement: | BR23 |
| Summary: | Data custodian can set the access rules for data. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H23, BASELINE REQUIREMENT 23**

*BR24:*

| | |
|---|---|
| Requirement: | BR24 |
| Summary: | Data custodian can review the data’s metadata. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H24, BASELINE REQUIREMENT 24

*BR25:*

| | |
|---|---|
| Requirement: | BR25 |
| Summary: | Data custodian can update the data’s metadata. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H25, BASELINE REQUIREMENT 25

*BR26:*

| | |
|---|---|
| Requirement: | BR26 |
| Summary: | Data custodian can review the access rules for distilled data. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H26, BASELINE REQUIREMENT 26

*BR27:*

| | |
|---|---|
| Requirement: | BR27 |
| Summary: | Data custodian can set the access rules for distilled data. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H27, BASELINE REQUIREMENT 27

*BR28:*

| | |
|---|---|
| Requirement: | BR28 |
| Summary: | Data custodian can review the distilled data’s metadata. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H28, BASELINE REQUIREMENT 28

*BR29:*

| | |
|---|---|
| Requirement: | BR29 |
| Summary: | Data custodian can update the distilled data’s metadata. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H29, BASELINE REQUIREMENT 29

*BR30:*

| | |
|---|---|
| Requirement: | BR30 |
| Summary: | Data custodian can review the access rules for processed data. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H30, BASELINE REQUIREMENT 30**

*BR31:*

| | |
|---|---|
| Requirement: | BR31 |
| Summary: | Data custodian can set the access rules for processed data. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H31, BASELINE REQUIREMENT 31**

*BR32:*

| | |
|---|---|
| Requirement: | BR32 |
| Summary: | Data custodian can review the processed data’s metadata. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H32, BASELINE REQUIREMENT 32**

*BR33:*

| | |
|---|---|
| Requirement: | BR33 |
| Summary: | Data custodian can update the processed data’s metadata. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H33, BASELINE REQUIREMENT 33**

*BR34:*

| | |
|---|---|
| Requirement: | BR34 |
| Summary: | Data custodian can review the access rules for actionable information. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H34, BASELINE REQUIREMENT 34**

*BR35:*

| | |
|---|---|
| Requirement: | BR35 |
| Summary: | Data custodian can set the access rules for actionable information. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

**Table H35, BASELINE REQUIREMENT 35**

*BR36:*

| | |
|---|---|
| Requirement: | BR36 |
| Summary: | Data custodian can review the actionable information's metadata. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H36, BASELINE REQUIREMENT 36

*BR37:*

| | |
|---|---|
| Requirement: | BR37 |
| Summary: | Data custodian can update the actionable information's metadata. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H37, BASELINE REQUIREMENT 37

*BR38:*

| | |
|---|---|
| Requirement: | BR38 |
| Summary: | Data subject has given consent for gathering, sharing and processing of his personal data that is ingested into Business Data Lake. |
| Priority level: | Low |
| Status: | Fully implemented at baseline state. |
| Comment: | No change required. |

Table H38, BASELINE REQUIREMENT 38

## Target Architecture Vision's Functional Requirements

*TR01:*

| | |
|---|---|
| Requirement: | TR01 |
| Summary: | Data custodian has the capability to manage personal data specifically. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Key requirement for solving the compliance of BDL with GDPR. |

Table H39, TARGET REQUIREMENT 01

*TR02:*

| | |
|---|---|
| Requirement: | TR02 |
| Summary: | Data custodian has the capability to manage Data subjects. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Key requirement for solving the compliance of BDL with GDPR. |

Table H40, TARGET REQUIREMENT 02

*TR03:*

| | |
|---|---|
| Requirement: | TR03 |
| Summary: | Data custodian's access to management of personal data and Data subjects is limited by access rights. |
| Priority level: | Normal |
| Status: | Envisioned |
| Comment: | Key requirement for solving the compliance of BDL with GDPR. |

Table H41, TARGET REQUIREMENT 03

*TR04:*

| | |
|---|---|
| Requirement: | TR04 |
| Summary: | Data custodian has the capability to manage requests from Data subject that consist of Request for a copy of personal data, Request for erasure and Withdrawal of consent. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Adds the functionality of managing different Data subject's requests to the system. |

Table H42, TARGET REQUIREMENT 04

*TR05:*

| | |
|---|---|
| Requirement: | TR05 |
| Summary: | Data custodian has the capability to report to Data Protection Authority. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Adds the functionality of reporting to DPA to the system. |

Table H43, TARGET REQUIREMENT 05

*TR06:*

| | |
|---|---|
| Requirement: | TR06 |
| Summary: | Data custodian has the capability to search through personal data. |
| Priority level: | Normal |
| Status: | Envisioned |
| Comment: | Possibly redundant with TR01 |

Table H44, TARGET REQUIREMENT 06

*TR07:*

| | |
|---|---|
| Requirement: | TR07 |
| Summary: | Data custodian has the capability to create a copy of personal data in an interchangeable format. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Relates to Right to receive a copy of personal data GDPR requirement. |

Table H45, TARGET REQUIREMENT 07

*TR08:*

| | |
|---|---|
| Requirement: | TR08 |
| Summary: | Data custodian has the capability to communicate with the Data subject that made a request, including sending files. |
| Priority level: | Normal |
| Status: | Envisioned |
| Comment: | Capability to inform Data subject of the status of his request and to send a file (copy of personal data). |

Table H46, TARGET REQUIREMENT 08

*TR09:*

| | |
|---|---|
| Requirement: | TR09 |
| Summary: | Data custodian has the capability to erase personal data. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Relates to Right to erasure GDPR requirement. |

Table H47, TARGET REQUIREMENT 09

*TR10:*

| | |
|---|---|
| Requirement: | TR10 |
| Summary: | Data custodian has the capability to create and manage filtering rule at ingestion stage. |
| Priority level: | Low |
| Status: | Envisioned |
| Comment: | Relates to Withdrawal of consent GDPR requirement. The functionality probably already exists at baseline state. |

Table H48, TARGET REQUIREMENT 10

*TR11:*

| | |
|---|---|
| Requirement: | TR11 |
| Summary: | Data custodian has the capability to create a standing search-and-erase query. |
| Priority level: | Low |
| Status: | Envisioned |
| Comment: | Relates to Withdrawal of consent GDPR requirement. The functionality probably already exists at baseline state. |

Table H49, TARGET REQUIREMENT 11

# Appendix I

## Business Transformation Risks and Mitigation Activities

| Risk ID | Risk | Preliminary Risk | | | Mitigation | Residual Risk | | |
|---|---|---|---|---|---|---|---|---|
| | | Effect | Freq. | Impact | | Effect | Freq. | Impact |
| R01 | Finalized version of GDPR differs from draft version, causing Target architecture to fail in solving all aspects of compliance with GDPR. | Catastrophic | Frequent | E | The process of GDPR adoption would have to be monitored closely, adjusting the Target architecture to changes as they are made. | Marginal | Occasional | M |
| R02 | Hadoop immutability does not allow for data to be physically erased, raising questions about validity of Right to erasure implementation. | Critical | Frequent | E | Acceptance of alternatives to physical erasure is expected to be defined in the process of GDPR finalization. A workaround involving file locks shall be considered if no other solution is available. | Marginal | Occasional | M |
| R03 | GDPR requirements are not correctly understood due to the lack of legal expertise in reading the law's text. | Critical | Likely | H | A number of articles about GDPR in professional magazines will be consulted in order to verify the correct understanding of requirements. | Negligible | Seldom | L |
| R04 | The results of this work are not accepted by Project sponsor, or the project is not completed in time, resulting in that Business Data Lake is not brought to compliance with GDPR. | Catastrophic | Occasional | H | Project sponsor is likely to have at least several parallel projects for solving the compliance issue with GDPR. External solutions can also be purchased. | Marginal | Unlikely | L |
| R05 | Planned interviews fail to be realized, resulting in incomplete information about baseline architecture and/or possible solutions. | Catastrophic | Seldom | H | Alternative sources of information such as documents, manuals and other correspondents will be used. | Negligible | Seldom | L |
| R06 | Project documentation is destroyed in a catastrophic event such as fire, earthquake, flood, equipment failure, etc. | Catastrophic | Seldom | H | All project documentation is replicated in real-time using Dropbox cloud storage service. | Negligible | Seldom | L |
| R07 | Stricter requirements on personal data management impair the system's | Critical | Occasional | H | Anonymization of data will not be considered as a viable solution. Business | Marginal | Seldom | L |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | business capability in an unacceptable way. | | | | capability will be considered as a crucial requirement in choosing a solution for personal data management. | | | |
| R08 | Lack of competence in the project group causes the project to result in failure. | Catastrophic | Occasional | H | External expertise will be acquired in areas where competence is lacking. | Marginal | Occasional | M |
| R09 | Inadequate file format for transferring personal data across systems is chosen. | Marginal | Seldom | L | Only standardized open source formats will be considered as candidates for interchangeable file format. | Negligible | Seldom | L |
| R10 | The project will not be completed in time or it will be done poorly because of limitations in manpower. | Critical | Occasional | H | The chosen methods for architecture development and project organization shall be followed zealously. Assigned deadlines will be honored to the best of project group's abilities. | Marginal | Seldom | L |

Table I1, Business Transformation Risks and Mitigation Activities

# Appendix J

## Additional requirements captured during design

*DR01:*

| | |
|---|---|
| Requirement: | DR01 |
| Summary: | Access to all data must be regulated through access control. This is especially important for personal data. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Prevents unauthorized access to data. |

Table J1, DESIGN REQUIREMENT 01

*DR02:*

| | |
|---|---|
| Requirement: | DR02 |
| Summary: | Unauthorized access to personal data must be prevented through the use of access control mechanisms. A weak spot in security is the ability of Data subjects to have access to their personal data. A special care must be taken in designing access control mechanisms and establishing authenticity of Data subjects. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Prevents personal data from being revealed to or destroyed by an entity other than the Data subject the data relates to. |

Table J2, DESIGN REQUIREMENT 02

*DR03:*

| | |
|---|---|
| Requirement: | DR03 |
| Summary: | Communication and information exchange with a Data subject must be established through a secure connection that guarantees confidentiality, integrity and availability of communication, but also its authenticity, accountability and non-repudiation. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Related to DR02 |

Table J3, DESIGN REQUIREMENT 03

*DR04:*

| | |
|---|---|
| Requirement: | DR04 |
| Summary: | The state when a personal data is ingested and stored without identification as such is technically a violation of GDPR and should be avoided as such at all costs. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Key requirement that enables personal data management. Requires cooperation with data sources. Perhaps impossible to guarantee 100%. |

Table J4, DESIGN REQUIREMENT 04

*DR05:*

| | |
|---|---|
| Requirement: | DR05 |
| Summary: | Consistent and current records of personal data in storage must be kept. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Management of personal data must be performed consistently and immediately in order to limit the possibility of GDPR violation. |

Table J5, DESIGN REQUIREMENT 05

*DR06:*

| | |
|---|---|
| Requirement: | DR06 |
| Summary: | A search for personal data should not require a full search of the complete repository for yielding all personal data related to a single Data subject. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Related to DR05 in that it enables immediate management of personal data. |

Table J6, DESIGN REQUIREMENT 06

*DR07:*

| | |
|---|---|
| Requirement: | DR07 |
| Summary: | All instances (or versions) of personal data that physically exist on storage have to be registered in the personal metadata repository. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Related to DR05 in that it enables immediate management of personal data. Important in the light of Hadoop fault tolerance and reliability that causes immutability as a side effect. |

Table J7, DESIGN REQUIREMENT 07

*DR08:*

| | |
|---|---|
| Requirement: | DR08 |
| Summary: | A record of made requests, received replies and received personal information should be available to Data subject. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Enables Data subject to manage and follow-up his requests regarding Data subject's personal data. |

Table J8, DESIGN REQUIREMENT 08

*DR09:*

| | |
|---|---|
| Requirement: | DR09 |
| Summary: | New data coming out from processing sandboxes should be checked for personal data before being released to storage. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Through processing and aggregation, even data that previously was not classified as personal can become that. |

Table J9, DESIGN REQUIREMENT 09

*DR10:*

| | |
|---|---|
| Requirement: | DR10 |
| Summary: | Personal metadata repository should provide key-value pairs that connect the Data subject with the locations of his personal data in Business Data Lake. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Data subject is linked to all of his personal data through Personal metadata repository. |

Table J10, DESIGN REQUIREMENT 10

*DR11:*

| | |
|---|---|
| Requirement: | DR11 |
| Summary: | A copy of personal data should provide an identical copy of personal data to greatest possible extent in regard to transformation of file format. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Transformation of file format may have side effects on the information the file contains. |

Table J11, DESIGN REQUIREMENT 11

*DR12:*

| | |
|---|---|
| Requirement: | DR12 |
| Summary: | Data Protection Officer should be able to manage and keep record of requests from Data subjects. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Records of requests and their management should be kept. |

Table J12, DESIGN REQUIREMENT 12

*DR13:*

| | |
|---|---|
| Requirement: | DR13 |
| Summary: | Information exchange between Data subjects and Data Protection Officer should be separated from the main system. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Security risks for the main system where data and personal data are kept should be limited by not allowing direct access for Data subjects. |

Table J13, DESIGN REQUIREMENT 13

*DR14:*

| | |
|---|---|
| Requirement: | DR14 |
| Summary: | Reliability and fault tolerance properties of Hadoop should not be impaired by management of personal data. |
| Priority level: | High |
| Status: | Envisioned |
| Comment: | Technical implementation of definite erasure of personal data may prove challenging. |

Table J14, DESIGN REQUIREMENT 14

# Appendix K

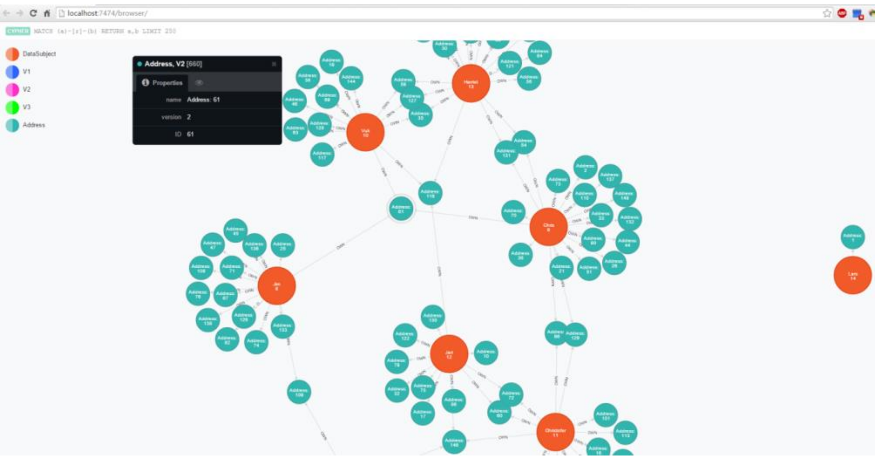
## Prototype construction code in Cypher

```
WITH ["Vuk","Christofer","Jari","Harriet","Lars","Erika","Husamedin","Stefan","Jim","Chris"] AS names
FOREACH (r IN range(1,100) | CREATE (:DataSubject {id:r, name:names[r % size(names)]+" "+r}));

WITH ["Address:"] AS names
FOREACH (r in range(1,50) | CREATE (:Address:V1{ID:r, version:1, name:names[r % size(names)]+" "+r}));

WITH ["Address:"] AS names
FOREACH (r in range(51,100) | CREATE (:Address:V2 {ID:r, version:2, name:names[r % size(names)]+" "+r}));

WITH ["Address:"] AS names
foreach (r in range(101,150) | create (:Address:V3 {ID:r, version:3, name:names[r % size(names)]+" "+r}));
MATCH (d:DataSubject),(a:Address)
WITH d,a
SKIP 1000
LIMIT 5000
WHERE rand() < 0.1
WITH d,a
LIMIT 100
MERGE (d)-[:OWN]->(a);
```

## Screenshots of graph model prototype in Neo4j

```
MATCH (a)-[r]-(b) RETURN a,b
```

// Returns nodes and their connections. Only a part of the graph is shown on Figure K1 due to the screen size limitation.



localhost:7474/browser/

CYPHER MATCH (a)-[r]-(b) RETURN a,b LIMIT 250

Address, V2 [660]

Properties

name Address: 61

version 2

ID 61

Figure K1, A partial view of the constructed graph database

# Appendix L

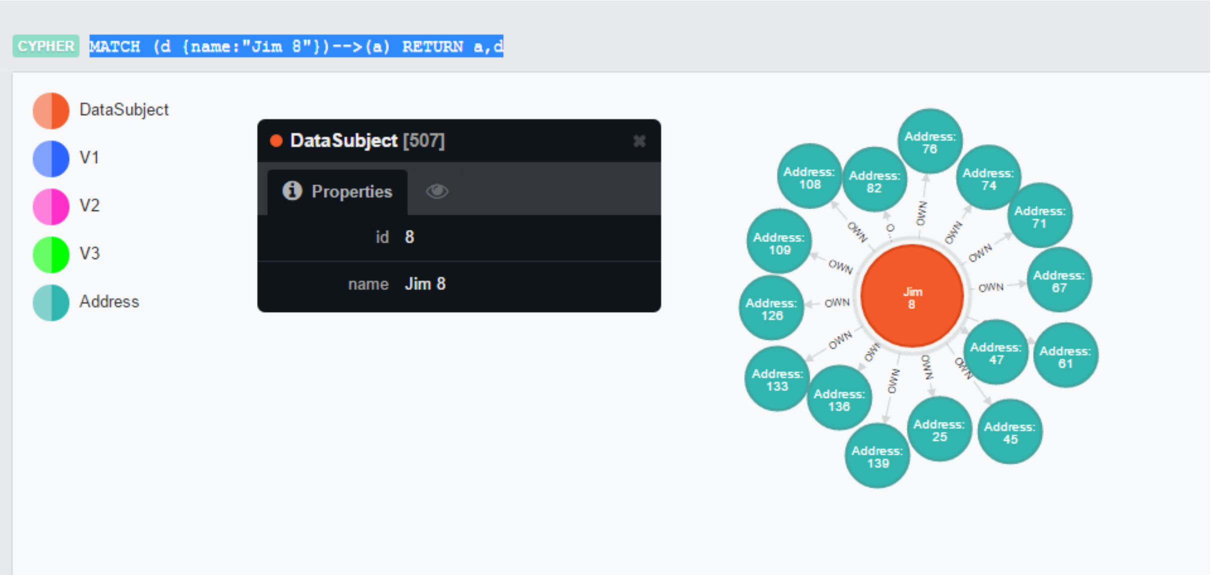
## Query examples

**Example 1:**

Find all data related to “Jim 8”:

MATCH (d {name:"Jim 8"})-->(a) RETURN a,d

//Figure L1 shows the visualization of the result of the query



```
CYPHER MATCH (d {name:"Jim 8"})-->(a) RETURN a,d
```

DataSubject [507]

Properties

| | |
|---|---|
| id | 8 |
| name | Jim 8 |

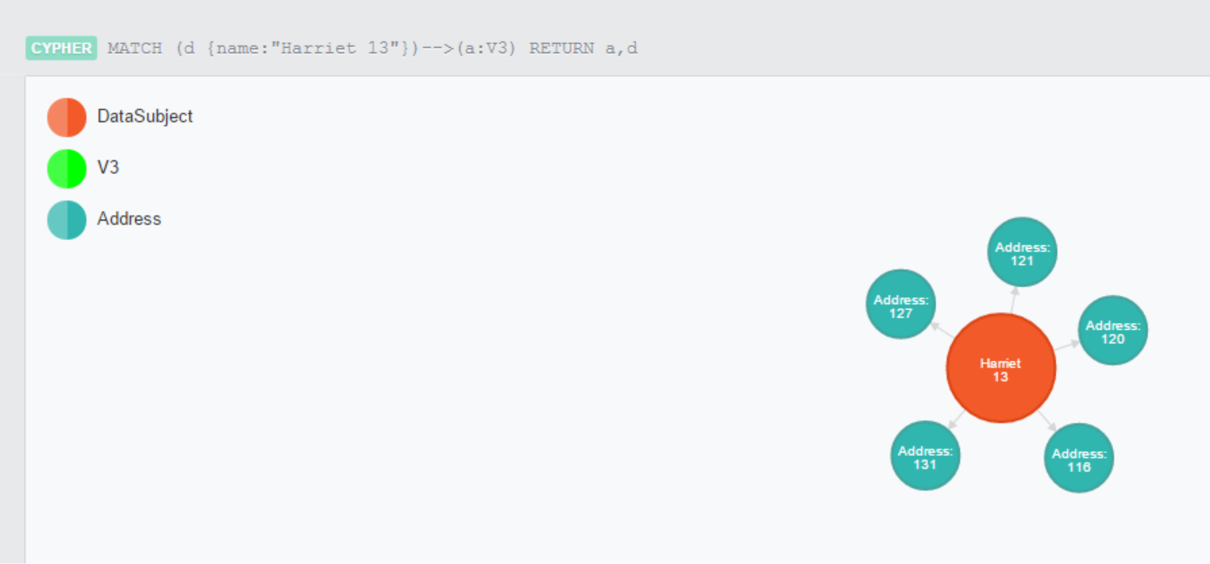
Figure L1, Query example - find all data related to "Jim 8" Data subject

**Example 2:**

Find all the latest (third for all data in this prototype) version of data related to “Harriet 13”:

MATCH (d {name:"Harriet 13"})-->(a:V3) RETURN a,d

//Figure L2 shows the visualization of the result of the query



```
CYPHER MATCH (d {name:"Harriet 13"})-->(a:V3) RETURN a,d
```

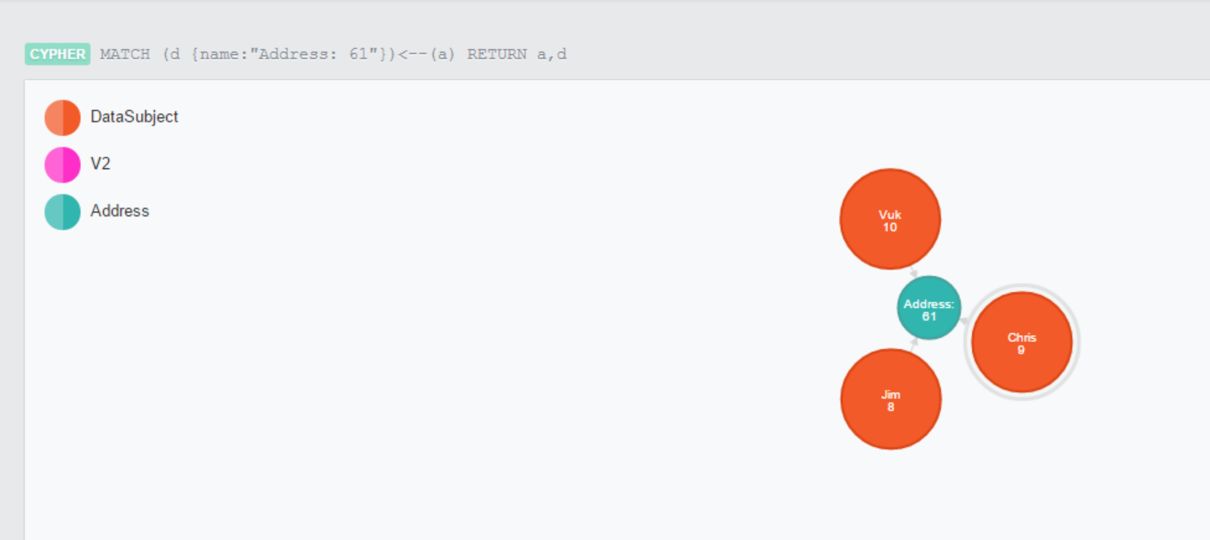
Figur L2, Query example - find all data of version 3 that relates to "Harriet 13" Data subject

**Example 3:**

Find all Data subject that file on location “Address: 61” refers to:

MATCH (d {name:"Address: 61"})<--(a) RETURN a,d

//Figure L3 shows the visualization of the result of the query



```
CYPHER MATCH (d {name:"Address: 61"})<--(a) RETURN a,d
```

Figure L3, Query example - find all Data subjects connected to the file at location “Address: 61”