Haaga-Helia
University of Applied Sciences

Creating value using Big Data Governance approach

Pavel Gavrilov

Master's Thesis
Degree Programme in
Business Administration
2020

Haaga-Helia
University of Applied Sciences

Abstract

28.12.2019

Author
Pavel Gavrilov

Degree programme
Business Administration, Master's Degree

Report/thesis title
Creating value using Big Data Governance approach

Number of pages and appendix pages
69 + 4

This research is a case study which covers implementing Big Data Governance approach in small-sized organization. The research is action research.

Case organization provides web-application based service for real estate operations and has huge amount of data related to real estate properties and their usage. Main purpose of this research is to utilize this data using modern Big Data solutions for creating new value and insights. Another purpose of this research is to show how Data Governance is applicable to Big Data.

Theoretical framework of this research is based on Data Management Framework and modern Big Data cloud solutions.

Implementation of this case study consists of two pilot projects with different technical approaches forming proof-of-concept for case organization. First pilot project used data about real estate properties. As a result of this pilot project, value to case organization's customers was provided in form of new visual reports. Second pilot project used data from web applications usage logs. As a result, this pilot project generated new insights, which enriched further software development of case organization.

The research proved that modern Big Data solutions allow to process huge amounts of data in small-sized organization without previous experience in this area. Also, the research indicates importance of Data Governance especially in manipulating Big Data.

In the end of this research theoretical framework is validated against the experience gained from implementation phase. This comparison highlighted similarities of pilot projects despite differences in technical approaches. Finally, based on these similarities Data Governance Framework for Data Processing is formed.

Keywords
Big Data, Data Governance, Data Management Framework

Table of contents

Index of tables

Table of figures

1 Introduction

"Ignoring the big data revolution is a very risky approach for any small business to take." (Marr 2019)

As a term "Big Data" has been here for a while. However Big Data Governance is not something that arises often to the news topics. As Washington (2019) points out, Big Data can be valuable only if Data Governance is handled correctly. One of purposes of this thesis is to test on practice, how Big Data Governance can be applied in small-sized organization. Another purpose is to develop valuable Big Data project for case organization. Below in this chapter objectives, case, scope and terminology are described in detail.

1.1 Objectives and Research Questions

The main purpose of this thesis is to examine possibilities of well governed Big Data to create value for small business. Primary objective from perspective of case organization is to find and to test modern Big Data solutions which could be utilized in development. Here "test" means attempt to create new value for case organization or its customers. Primary objective from perspective of personal evaluation is to broaden understanding of Data Governance, Big Data, their relation and effectiveness in small-sized organization. Putted in form of research questions these objectives are:

- Can small-sized organization create value from Big Data without previous experience in that area?
- Can Big Data Governance be handled with respect in situations, where human resources are strictly limited?

1.2 Background and Case Overview

Case organization of this thesis continuously develops its own service for all real estate users such as residents, caretakers, real estate managers, owners and mutual funds. Case organization provide the service in form of SaaS (Software as a Service) and it is served via web-based applications, so it is available from any place and any device with Internet connection. There are tens of thousands real estate properties, which data is managed using case organization service and tens of thousands of daily users. Therefore, case organization has vast amount of data. Although case organization has numerous data processing operations, modern Big Data tools can possibly create new insights of existing data. These insights can be valuable directly to end users of case organization's service or support further development of the service.

1.3 Scope

Scope of this thesis is limited to one case organization and developing proof-of-concept based on two pilot projects. First reason for taking two pilot projects instead of only one was in better validation of suitability of selected Data Management Framework. Second reason was in testing different technical approaches of data processing. More pilot projects or more case organizations could bring deeper understanding of both aspects, but limited time restricted scope of this thesis.

2 Methodology

This chapter describes used research strategy and data collection methods.

The main goal of this study is to create new customer value from existing data by implementing new manner of processing and analyzing data in target organization. Therefore, qualitative approach is reasonable choice for the study. According to Denzin & Lincoln (2000, 14) qualitative research provides multiple tools to collect empirical material like questionnaires, interviews and documents analysis. Also, McLeod (1994) points out that study on data which is not easily calculable is always qualitative.

2.1 Research Strategy

Nature of the research question narrowed selection of research strategy to two possible choices: case study and action research. They both are nominally suitable for the case and both provide good variety of tools to find answer to the research question.

Case study is a common name for the set of studies which are generating insights on the real-life phenomena through in-depth analysis of the case or the cases (Saunders, Lewis & Thornhill 2016, 185.) Most suitable type of case study for the current research is single case study. Making deep inquiry on the target case could achieve awareness of successfulness of the development project and generate useful suggestions for similar projects in future inside and outside the target organization. However, case study does not suppose trying to affect progress of current project within a study not to mention actual planning of the project. As Saunders et al (2016, 186-187) points out, main target of all case studies is to observe and deductively create perspectives.

Action research as an implementation of case study defines more tight interaction with the case as researcher is also involved into planning and implementing actual case. Action research promote solutions using iterative process (Saunders et al 2016, 190.) This process consists of four steps that are repeated multiple times during the research. Each cycle of iteration brings new insights on the problem basing on knowledge gained during previous cycle (Saunders et al 2016, 191.) Visualization of the process (Figure 1) clarifies uniformity of research and basics of continuous development processes described in chapter 4.

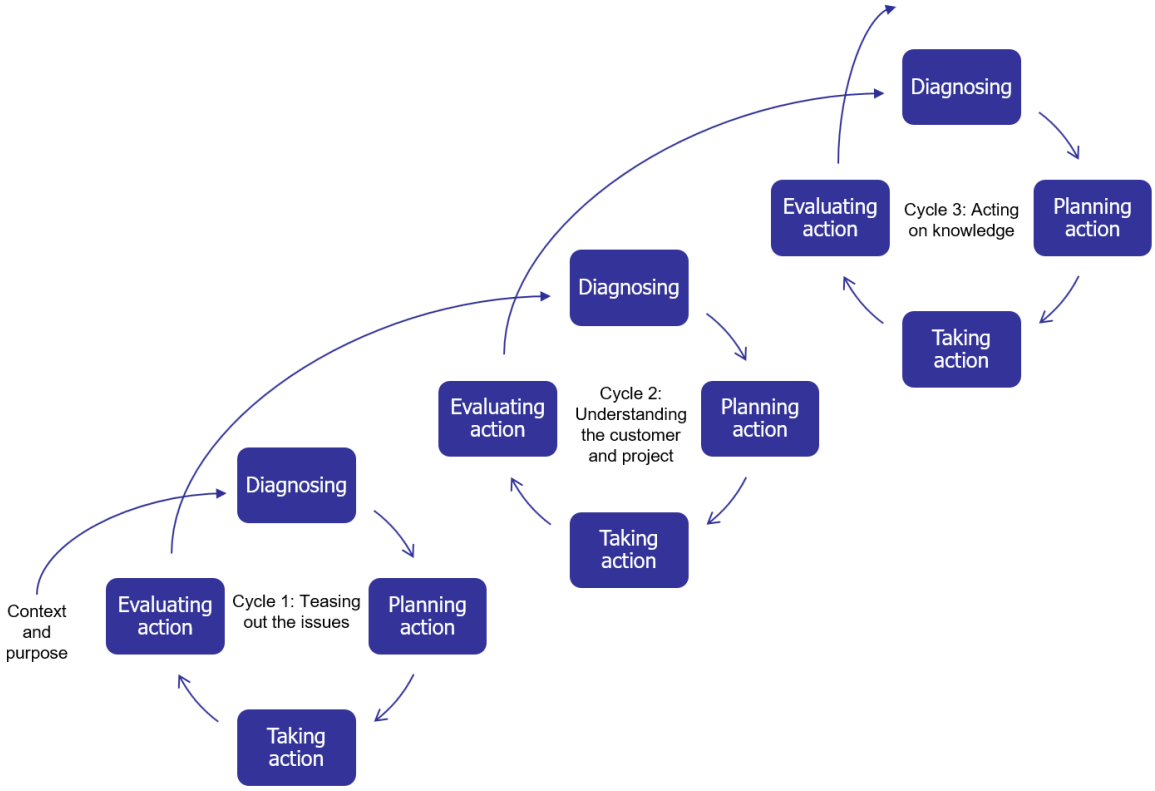


Figure 1. Action research process (Saunders et al 2016, 191)

Main goal of the action research is to learn thru taking actions on resolving real case within organization. Investigator is deeply involved in resolving process and his or her activity not only results on the research outcome but powerfully affects organizational development process related to the project. That participation form better case understanding and commonly generates valuable insights on the issue which hardly can be generated using any other research method. (Saunders et al 2016, 191-192).

Comparing different research methods suitability for the research question leaded to understanding that action research is more applicable. It provides better approach for resolving the research question by allowing taking active role in it. Furthermore, action research process supports continuous development and therefore empower improvement of the business case even after the research.

2.2 Data Collection Methods

Data source types can be divided in two base classes: primary data and secondary data. Primary data is any raw data collected from original sources by questionnaires, interviews, observations and other research-based data collection methods. Secondary data is previously collected and structured data results from other researches, which can be useful for current research. (Saunders et al 2016, 316-317).

The very first step in deciding appropriate methods of collecting data is understanding what data collecting results we are about to gain. That leads us to finding the best available source of data needs to be collected. Below table shows essential data collections for successful research result and the best available source for gaining the data:

Table 1. Data to be collected and data sources.

Data description	**Source description**
Big Data possibilities	Books Web publications
Available Big Data solutions	Web publications
Implementation possibilities	Books Software developers
Technical aspects of implementation	Software developers Web publications
Business aspects of implementation Target customer group	Business owners
Customer expectations Customer satisfaction	Target customers

Books and web publications mentioned in Table 1 are initial secondary data collected to gain overall understanding on the matter. There is a lot of researches and case studies on Big Data possibilities and in current case they are the best possible source for basis in terms of availability and usefulness. Also available solutions for Big Data implementation and their pros cons and options will be collected using mix of primary and secondary data from web publications because field of the subject is so dynamic that books may not contain up to date information and scope restrictions does not include option for testing them all.

As we can see in Table 1, there is three groups of people which are considered as data source: software developers, business owners and target customers. Two first groups are small, but the data available from them is highly important for the study. Because of that, it was urgent to select best available data collection method for each group.

The research interview is very powerful method of collecting primary data. It helps to get insights and deeper understanding of research objectives. Interview provides subjective

perspective and it's result highly depends on interaction between interviewer and interviewee(s). All that putted together makes choosing appropriate type of interview and good preparation for interview extremely important. (Saunders et al 2016, 388-390).

One of the most common typologies of interviews provides three types: structured interviews, semi-structured interviews and unstructured or so-called in-depth interviews. Structured interviews are very formal, they use questionnaires and are based on idea that no question or even voice tone of interviewer is changed between interviewees, so they fit well together with quantitative research method. Two other types of interviews are more suitable for qualitative researches. Semi-structured interviews usually contain some initial questions and main themes but structure of them can vary between interviewees as more topic context is gained during interviews. Unstructured interviews may not contain any predefined form or any relation to previously completed interviews. (Saunders et al 2016, 390-392).

Organizations own secondary data such as emails, documents, memos and database are very valuable source on its' own. In addition to that, secondary data can provide good basis for forming interviews. Because two main purposes of all interviews done for this research are collecting data and comparing expectations to the actual satisfaction, semi-structured interviews was chosen as a main interview type. In addition to that there were several occasional unstructured interviews as they are natural part of action research done by process development in team.

3 Big Data Governance

This chapter describes data governance in a nutshell and big data governance as a part of corporation strategy towards data driven development. It goes thru common practices of data governance and most popular frameworks of implementing data governance processes in companies. In the end of the chapter theoretical background is reflected against main goals of current research and most suitable big data governance framework is chosen.

3.1 Data Governance

Data Governance is a term for describing key principles of management available data in organizations in a way which helps to take most advantage of it and make the data valuable (DAMA International 2017, 67-68) Data Management Framework is shown below (Figure 2) and it shows key areas of it. Data Governance is placed to the center of the framework to bring up its importance and key role in all other areas.

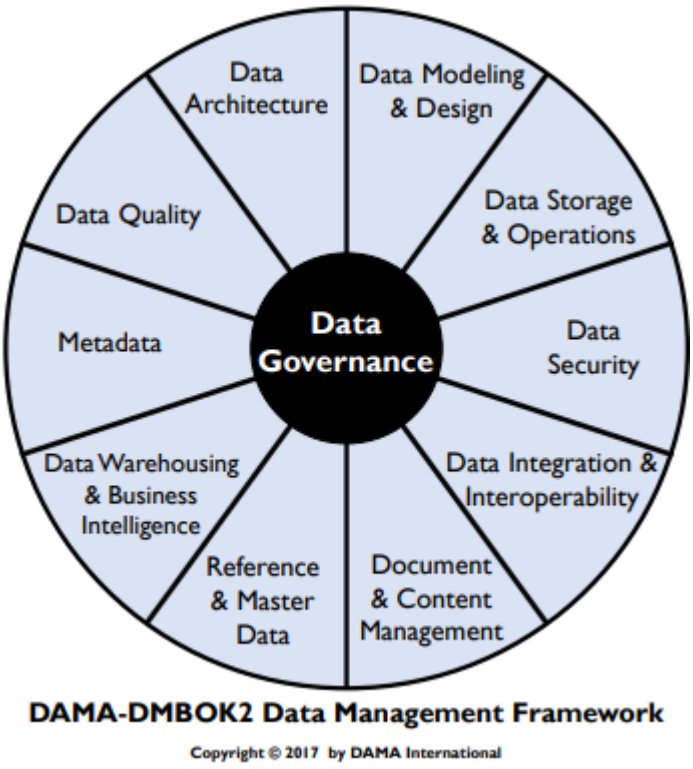


Figure 2. Data Management Framework (DAMA International 2017, 67)

Data Governance is not a single time project nor even an activity which can be accomplished by some individuals or department within organization. It is a concept for creating understanding on all levels of organization to concern data as a strategic asset. Data Governance gives guidelines to continually improve activities to maintain data on high level of strategical value. (DAMA International 2017, 68).

According to DAMA International (2017, 70) main drivers for Data Governance are reducing risks and improving processes. Reducing risks enclose all risks related to data security, privacy and compliance to local and global legislation while improving processes in this scope determine all activities that support organizations strategy to become more competitive using data governance principles (DAMA International 2017, 70-71.) Sarsfield (2009, 38-40) indicate, that main driver for most of Data Governance programs comes from c-level executives and their emergent need to make data-based decisions while amount of collected data grows exponentially. Also, Ladley (2012, 34) states that main benefits of Data Governance are efficiency improvement, business contributors increment and risk reduction. DAMA International (2017, 73) summarizes that successful implementing of data governance and considering data as an asset is possible only if it is in line with business strategy and its importance is understood across the organization.

3.1.1 Metadata

Most commonly Metadata is defined as “data about data”. Castanedo & Gidley (2017) divide Metadata in three types: 1) Technical metadata, which describes data from technical viewpoint like type, structure and size of data; 2) Operational metadata, which contains information about data quality, location, update cycles and lineage; 3) Business metadata, which contains definitions, that are understandable by end users of data and therefore describes business value of data. Another point of view on dividing Metadata is defined by Pomerantz (2015, 17-18), he also divides Metadata in three types and they are: 1) Descriptive metadata, data that helps discover resources and information, such as catalogs; 2) Administrative metadata, data that helps usage of information, containing also sub categories like Structural metadata and Preservation metadata; 3) Use metadata, data that contains information about resource usages, like file download logs. In comparing with three types described previously, administrative metadata with its subcategories is very close to technical metadata and operational metadata, while descriptive metadata and use metadata are closer to business metadata. All that putted together gives understanding, that metadata has no singe all-purpose approach, but rather approaches and methods for generating and maintaining metadata depend on the desirable result and usage cases of metadata. Both Castanedo & Gidley (2017) and Pomerantz (2015, 32-33) argue, that business glossary or in another words “controlled vocabulary” helps maintaining metadata independently from type of metadata. Controlled vocabulary provides limited amount of definitions and key phrases do describe metadata and therefore keeps metadata more structured.

Metadata needs to be managed as it provides users with information about the data in general. As T. Nero (2018) points out, “A metadata management strategy is central in ensuring

that data is well interpreted and can be leveraged to bring results. Such metadata management strategies include collection, storage, processing, and cleaning." Well-organized Big Data management is not possible without metadata management. Accordingly, an understanding of metadata management processes plays an important role in organizing Big Data. Metadata management is central to data management. Without metadata management, the data management of any company cannot be organized. (Nero 2018).

The concept of Data Governance contains metadata as a component. Metadata management allows analytics to understand the current business environment and predict its development better, revealing the features of interaction with it. Some companies use the same tools for managing metadata as for managing Big Data. Metadata can also be accessed by users. (Nero 2018).

Metadata improves communication between heterogeneous information systems. Metadata management makes it possible to apply more secure management methods in the future, to ensure the confidentiality of the information and to prevent data leakage. The importance of metadata in the contemporary world continues to grow since raw data must be supplemented with metadata in order to fully unlock the potential of new technologies. The expansion of the digital universe means an increase in the amount of useful data in it. (Schmarzo 2018).

3.1.2 Data Architecture

Data Architecture is a fundamental discipline of Data Management, which covers activities to describing the current state of data and data flows within the organization and design desirable state of it. Key outcome of data architecture, or in another words "Data Architecture Artifacts" is abstract enterprise data model understandable by organizations management, containing Metadata, data relationships and business rules including data flow rules. (DAMA International 2017, 98).

Enterprise Data Model (EDM) is high-level abstract model providing organization's comprehension of enterprise-wide data model, logical subject area data models, project- and application-related data models, their vertical and horizontal relations and business rules. Vertical relations describe relations from enterprise-wide data model to application-related data models while horizontal relations describe relations between models in the same level of abstraction. (DAMA International 2017, 105).

As can be found from Zachman Framework (Zachman 2008) during the definition of the target state, the Data Architecture breaks a subject down to the atomic level and then

builds it back up to the desired form. The database architect breaks the subject down by going through 3 traditional architectural processes:

- Conceptual - represents all business entities.
- Logical - represents the logic of how entities are related.
- Physical - the realization of the data mechanisms for a specific type of functionality.

Below is adopted Zachman Framework to the roles of stakeholders in organization merged with responsibilities from DAMA International (2017).

Table 2. Adopted from Zachman Framework

Layer	View	Data (What)	Stakeholder
Contextual	Scope	List of things and architectural standards important to the business	Planner
Conceptual	Business Model	Sematic model or Conceptual Data Model	Owner
Logical	System Model	Logical Data Model	Designer
Physical	Technology Model	Physical Data Model	Builder
Detailed	Representations	Actual databases	Subcontractor

Data architecture is used to manage Big Data. Big Data architecture is a system used to receive and process Big Data so that it can be analyzed for commercial purposes (Alley 2019.) The concept of Big Data architecture involves the processing and analysis of data that is too large for traditional data systems. It can be determined when companies are moving from data management to Big Data management basing on the size of the company – each business has its own threshold. Big Data architecture means that advanced analytics extract useful information from data. (Big Data Architectures 2018).

Most Big Data architectures include some of the following components, or all of these components at the same time: data storage, batch processing, real-time message ingestion, stream processing, analytical data store, analysis and reporting, and orchestration. Thanks to the big data architecture, predictive analytics, and machine learning can also be implemented. A well-designed big data architecture can help save the company's money and predict future trends in such a way that the business will only make the right decisions. (Alley 2019).

3.1.3 Data Modelling and Design

Data modelling is a process of discovering data requirements and documenting these requirements in visual form called data model, which is key deliverable of this process. As a form of Metadata, data models accomplish multiple key goals of data management: they provide data vocabulary; they keep record about data systems and data; they support communication about data; they are significant tool for maintaining, reusing and integration of data. Data models can be divided by the level of abstraction to three main types: Conceptual Data Model (CDM), Logical Data Model (LDM) and Physical Data Model (PDM). (DAMA International 2017, 124-125).

Conceptual data models represent key business entities, data requirements and relations between them (DAMA International 2017, 145.) Logical data models are more detailed representations with attributes of business entities described in CDM, but still abstract and do not cover technical implementations (DAMA International 2017, 146.) Physical data models go deeper in technical implementations of LDM and contain more information of actual implementation of data entities and their relations (DAMA International 2017, 148.)

Approach to build data model is going from up to down i.e. from CDM to LDM to PDM is called forward engineering and it is used to build data models for new applications (DAMA International 2017, 153.) Another approach, called reverse engineering, backward direction from PDM to LDM to CDM and suits for building data models based on existing data relations (DAMA International 2017, 158.)

According to Simsion et. al. (2012, 151-152), data modeling in practice is most reminiscent of design. Big Data modeling is necessary because huge amounts of data must be in order to be usable. Using data models, the business can organize and store data, including Big Data. The models and environments for storing data provide the following benefits for Big Data: increasing productivity, lowering costs, improving user experience and efficiency, reducing the number of errors in calculations and improving the quality of calculations. Thus, Big Data systems need high-quality data modeling techniques (Zhang 2018). There are various methods for modeling Big Data. The models used for this can be divided into relational and non-relational database models (Huda et. al. 2015).

3.1.4 Data Storage and Operations

Data Storage and Operations can be divided in two main sub-activities: database support and database technology support. First one includes planning, implementing, monitoring

and maintaining database environments, databases and data within them. Second one focuses on determining technical requirements, planning technical solutions and architecture, implementing it, managing and monitoring implemented technology. Because organization's data is crucial for business continuity, Data Storage and Operations play vital role and is very important for almost any organization. (DAMA International 2017, 169-171).

Initial step of database technology management is understanding business requirements, available tools and technologies discovering and combining these two aspects for selecting best possible technology suitable for business needs. Often it is reasonable to start evaluating technology with proof-of-concept (POC) pilot project to get better understanding about pros and cons of selected technology before implementing it widely across the organization. After implementation database technology needs to be managed and monitored for better support of application development, ensuring backup and recovery possibilities. In addition to that, monitoring for technology updates and discovering new technologies and their possibly benefits for business is also important. (DAMA International 2017, 194-195).

There are several problems associated with data storage and operations. First, this is a data security issue. Using new technologies for storing and working with data, users may be worried about their data. This is true, for example, in the case of cloud technology. (Zhou & Huang 2012, 1).

The problem of data storage does not have a universal solution. Before choosing where to store its structured and unstructured data, the company must understand how much available data it has and for what purpose it needs to be stored. Awareness of businesses regarding their data implies knowing this data, storing all the data including unstructured, establishing their own data storage policies, choosing solutions that match company data, maintaining data security, etc. (Lonoff Schiff 2013).

3.1.5 Reference and Master Data

Reference and Master Data management is a process of simplifying sharing most important data within an organization in a way that provides common understanding that this core data is an asset of whole organization. Reference Data is one of three subsets of Master Data, other two are Enterprise Structure Data and Transaction Structure Data. Key outputs of Reference and Master Data management are data models with requirements and integration patterns; reliable and reusable reference and master data. (DAMA International 2017, 348).

Reference Data classifies other data and contains codes (statuses or types), descriptions and mappings usually formed in some sort of attribute list that supports data standardization and verifying that data values are consistent across all organization units and functions (DAMA International 2017, 353.) Master Data contains key organizational real word entities representing them in data values trying to consolidate all data representing same entities across organization and minimize duplicate records of same entities (DAMA International 2017, 358.)

Master data may include reference data and all other data types that the organization permits to share. Thus, reference data is one of many types of master data. However, not every reference data can be considered as master data. The value of reference data is data standardization, validation, and data quality. The value of master data is data quality, reduced costs, operational efficiencies, and streamlining data management and governance. (Spacey 2016).

Master data is usually used by several IT systems and business processes, so standardizing this type of data format is critical to the success of system integration. As for reference data, “Organizations often create internal reference data to characterize or standardize their own information. Reference data sets are also defined by external groups, such as government or regulatory bodies, to be used by multiple organizations” (McGilvray & Thomas 2008, 1-2.)

3.1.6 Data Integration and Interoperability

Data Integration and Interoperability (DII) focuses on integration of different data sources, moving and consolidation of data to simplify, lower costs and improve effectiveness of data management and usage within different solutions and by different data consumers. DII activities are depended on many other areas of data management: Data Governance, Data Architecture, Data Security, Metadata, Data Storage and Operations, Data Modelling and Design. Data Integration and Interoperability activities support Business Intelligence and Master Data Management. (DAMA International 2017, 270-272).

Key process of all DII activities is Extract, Transform, and Load (ETL). The first step of this process contains picking up required data from source systems (Extract) either by scheduled batch processing to minimize resource usage and capture full data set at pickup time or by near-real-time and event-driven asynchronous process to capture data changes by schedule or trigger-event with smaller intervals or by real-time synchronous process to capture data changes at the moment they happens. The second step is making data suitable for target store (Transform) including format, structure, sematic, order changes in

data and duplicate removing from it. The last step of this process is publishing data to the target system (Load) for further processing and using by customers. (DAMA International 2017, 273-274).

In the case when the systems are compatible, they can not only exchange information but also interpret the data that enter them, presenting them in the form in which they were received. In other words, compatible systems speak the same language, while the integration process is more like talking in different languages. (Roberts 2017).

3.1.7 Data Warehousing and Business Intelligence

Data Warehousing (DW) focuses on extracting, transforming, loading and storing data in a way that makes is available for different analytic tools to support workflow, operational management and predictive analysis. Primary meaning of Business Intelligence (BI) is data analysis and discovering insights from data for making better decisions, taking data as competitive advantage and based on it fulfil business objectives and improve organizational success. Secondary meaning of BI are technologies and tools that support this kind of activities. Consequently, primary goal of Data Warehousing is supporting Business Intelligence activities and together they provide transforming data to a valuable asset and competitive advantage of organization. (DAMA International 2017, 383-384).

Developing DW/BI processes and activities starts with understanding requirements e.g. defining business goals and prioritizing desirable outputs. Next step is planning maintainable architecture from conceptual architecture models and prototyping providing understanding on possible technical solutions. After architecture is established developing of data warehouse and data marts can be started. Data mart is a subset of data warehouse, that focuses on a single business area like invoicing or stock. Other key activities of DW/BI are implementing BI portfolio and maintaining data products as business intelligence outputs needs to be made available for right business units and data in data warehouse needs continuous maintain activities to be correct and up to date. (DAMA International 2017, 395-399).

Currently, business intelligence and data warehousing are no longer synonymous. It was almost impossible to do business analytics without data storage before. Now there is the following alternative – instant BI in a data lake and automated data warehouses with ELT. The relationship between business intelligence and data warehousing was inconvenient and had disadvantages. However, currently, there are ways to overcome possible problems in this area. Effective business intelligence can be achieved even without data storage. (BI and Data Warehousing: Do You Need a Data Warehouse Anymore? 2019).

3.1.8 Document and Content Management

Document and Content Management is a process which main purpose is establishing activities for linking unstructured and even non-electronic data and information to structured data. Another key goal of Document and Content Management is to ensure security, availability, quality, lifecycle management and compliance of this data. These goals and purposes require architecture, governance and high-quality Metadata. (DAMA International 2017, 304).

One of key activities of Document and Content Management is building and managing controlled vocabularies, a type of Reference Data, which helps organizing and indexing various content and makes available for searching and browsing (DAMA International 2017, 309.) Another crucial activity is planning and managing lifecycle of content containing policies and rules for storing, retrieving, retention and destruction of documents according with regulations and organizational requirements (DAMA International 2017, 323.)

In some sources, document management and content management are considered not as synonymous concepts, but as processes that differ significantly from each other. Content management is carried out by a centralized center in which documents and other information are stored, which, if necessary, can be distributed. Document management takes place thanks to a computer system designed to search and store electronic documents. Document management is related to a broader area of content management. Content management is more suitable for managing unstructured data, while document management is usually being used for managing structured data. (Templafy 2019).

Interoperability is the ability to transfer data in such a way that the user does not even realize the unique characteristics of the systems between which the information exchange takes place. In the context of information technology, interoperability means that various devices, operating systems, and applications can work together efficiently without preconfiguration. The benefits of integrated solutions can also be described as follows: “Integrated solutions achieve superior interconnectivity, service delivery, and information access by offering the means to bring content and business logic together, regardless of the source, into one seamless user interface with single sign-on and authentication”. In other words, integrated solutions create infinite possibilities. (An Introduction to Integration and Interoperability 2003)

3.1.9 Data Security

Data Security main goals are protecting information assets from unauthorized access, ensuring authorized access and verifying alignment with regulations, agreements and privacy requirements. These goals can be accomplished by identifying security requirements and current risks; defining policies and standards and implementing continuous audit procedures. (DAMA International 2017, 217-219).

In addition to self-evident truth that security is valuable itself, data security has many other powerful business drivers. When ignored, data security risks can lead to failed compliance with governance regulations and contractual obligations. Data security is also very important because of customer confidence and organization reputation. Therefore, data security has high impact on planning risk reduction and ensuring business growth. (DAMA International 2017, 220).

Data Security activities starts with identifying organizational business requirements and relevant regulatory requirements. Good practice is keeping centralized set of high-level enterprise security policies and populate it to more specialized department and application policies. Another key activity is implementing data security controls and procedures. It contains defining data access rules, monitoring security risks mitigation, auditing policies fulfillment and reporting possible breaches and improvement suggestions. (DAMA International 2017, 247-248).

The amount of information used by businesses is increasing year by year. Most of the data is kept in unstructured form. As businesses want to capitalize on this information, make it useful and manage it, Big Data technologies are being used more and more often. However, in addition to capabilities, large amounts of information contain potential threats to examine and analyze data sets (Tawalbeh & Saldamli 2019.) Large data warehouses pose new security concerns, especially for the direct use of data. One of the possible solutions to the data security problem may be the placement of data controls not on sites and not in applications where data is stored, but around the data itself (Tankard 2012, 5.)

Big Data security adds value to the data. The quality and security of Big Data are their main features that make data more meaningful. Therefore, before operating the data, it is important to ensure the safety and quality of Big Data. It is also worth noting that the safety and quality of Big Data depend on the overall context of the functioning of the technology. (Talhaa et. al. 2019, 916-917).

User data must be protected because it also contains confidential information. So, a data leak on social networks can lead to serious negative consequences. In general, any business should build a trusting relationship with the audience, and this is possible only in cases where the company cares about the security of user data.

3.1.10 Data Quality

Data Quality depend on and affect all other Data Management disciplines because high-quality data is the main goal of any data management activity and only high-quality data can be valuable for an organization. Therefore, data quality management needs to be cross functional and its importance should be appreciated by all teams and at all levels within the organization. Data Quality is not a standalone project but rather continuous program that includes projects for improving data quality, maintenance and trainings to support data quality at all stages of data lifecycle. (DAMA International 2017, 450).

Evaluation and improvement of data quality are possible only if there is a compromise between the quality and security systems of big data. Data quality is very important for business, because thanks to big data you can get authentic information for decision making (Cai & Zhu 2015.) Poor quality data can cause serious damage to the business at several levels, cause additional costs, loss of opportunities. Poor quality data also necessitates the re-execution of some important business tasks. (Talhaa et. al. 2019, 918.)

Effective implementation of systems to improve data security and quality is hampered by the diversity and rapid generation of big data. Such systems require well-designed mechanisms and strategies. The mutual influence of data quality and security systems requires further study since data security can be an obstacle to data quality and vice versa. Mutual conflicts between these systems pose new challenges that require new solutions. (Talhaa et. Al. 2019, 919).

The initial step of Data Quality Management is to define requirements for high-quality data and build consensus between data consumers about data quality improvements. This include understanding of current state of data, pain points and business goals. As almost any organization has a lot of data, it is important to specify critical data for the organization and focus data quality improvements on it. In addition to Master Data, critical data may be defined by risks of low-quality data for organization’s reputation, finance, regulatory compliance, ongoing operations or business strategy. (DAMA International 2017, 473-475).

Data Quality techniques can be divided in two main categories: preventive and corrective. Preventive techniques focus on eliminating poor data quality root causes and contain multiple approaches: establishing controls of data entry, training procedures, rules definitions and enforcement, external data quality requirements, defining roles and responsibilities, validating data change control. Corrective techniques focus on fixing poor quality data by automated correction using rule-based standards, manual-directed correction for sensitive data that requires human oversight before applying rule-based correction and least desirable manual correction in cases which can't be resolved by any other method. As preventive techniques are cheaper and less risky, they must be included in corrective techniques to prevent recurrence of data quality issues. (DAMA International 2017, 486-487).

To set up a data quality management system, the business needs to identify errors and anomalies that affect quality. Thanks to this phase of the work, new goals regarding data quality are set.

3.1.11 Roles and responsibilities

Different areas of Data Management have different participants with several areas of responsibility. In Table 3 is shown full list of possible participants according to DAMA International (2017).

Table 3. Participants of Data Management areas

DM Area	Participants.
Data Governance	Steering Committees, CIO, CDO / Chief Data Stewards, Executive Data Stewards, Coordinating Data Stewards, Business Data Stewards, Data Governance Bodies, Compliance Team, DM Executives, Change Managers, Enterprise Data Architects, Project Management Office, Governance Bodies, Audit, Data Professionals
Data Architecture	Enterprise Data Architects, Data Modelers
Data Modeling and Design	Business Analysts, Data Modelers
Data Storage and Operations	Database Administrator, Data Architect
Data Security	Data Stewards, Information Security Team, Internal Auditors, Process Analysts
Data Integration and Interoperability	Data Architects, Business and Data Analysts, Data Modelers, Data Stewards, ETL / Service / Interface Developers, Project and Program Managers

Document and Content Management	Data steward, Data management professional, Records management staff, Content management staff, Web development staff, Librarians
Reference and Master Data	Data Analysts, Data Modelers, Data Stewards, Data Integrators, Data Architects, Data Quality Analysts
Data Warehousing and Business Intelligence	Sponsors & Product Owner, Architects and Analysts, DW/BI Specialists (BI Platform, Data Storage, Information Management), Project Management, Change Management
Metadata Management	Data Stewards, Project Managers, Data Architects, Business Analysts, System Analysts
Data Quality Management	CDO, Data Quality Analysts, Data Stewards, Data Owners, Data Analysts, Database Administrators, Data Professionals, DQ Managers, IT Operations, Data Integration Architects, Compliance Team

As we can notice from Table 3, the list of participants is colossal, and it contains over forty different job titles. For big corporations it might be possible and appropriate to have so big Data Management team, but in target organization there is in total less employees than DAMA International (2017) described possible members of Data Management team and available resources for Data Management program doubtless are smaller than all organizations employees. That awareness led me to grouping Data Management areas by participants and taking in count only those ones, who might be involved in more than two Data Management areas. The result of that grouping is shown in Table 4.

Table 4. Data Management areas of most involved participants

Participant	DM Areas
Data Architects	Data Architecture, Data Storage and Operations, Data Integration and Interoperability, Reference and Master Data, Data Warehousing and Business Intelligence, Metadata Management
Data Modelers	Data Architecture, Data Modeling and Design, Data Integration and Interoperability, Reference and Master Data
(Business / Process / Data / System) Analysts	Data Modeling and Design, Data Security, Data Integration and Interoperability, Reference and Master Data,

	Data Warehousing and Business Intelligence, Metadata Management, Data Quality Management
Data Stewards	Data Governance, Data Security, Data Integration and Interoperability, Document and Content Management, Reference and Master Data, Metadata Management, Data Quality Management

As we can see from Table 4, it covers all of Data Management areas and most of them are listed at least twice. That finding supports belief that Data Management program can be rolled out even in small organizations with limited resources and strengthen understanding that every step of that program must be accomplished in a collaborative way. Below are described roles of the most involved Data Management participants.

Data modelers and data architects have many common areas of responsibilities, they are Data Architecture, Reference and Master Data and Data Integration and Interoperability.

One of key roles in Data Governance is Data Steward. Data Steward is a commonly used naming to characterize role of a person or a team who is responsible for effective Data Governance processes. Main areas of responsibility are supervising core Metadata, recording rules and standards, maintaining data quality and accomplishing operational data governance programs. Data Steward teams often include people who are responsible for linking Data Governance activities with organizations business strategy, for example Data Owners and Enterprise Data Stewards and people who are responsible for technical aspects of implementing Data Governance like Database Administrators and Data Quality Specialists. (DAMA International 2017, 76-77).

3.2 Big Data

This chapter starts with common description of Big Data and its existence in all IT areas and thru that its existence in all areas of human being. After that importance of Data Governance in Big Data solutions is demonstrated. Finally, most popular up to date Big Data tools and solutions are compared to find best possible toolset for actual implementation.

3.2.1 What is Big Data

The growth of information, its role, and amounts in the modern world gives rise to new technologies that help to streamline existing information. In the information society, companies use Big Data for in-depth interaction with customers, for preventing threats and

fraud, and optimize their activities. A quick introduction to this ambiguous technology will be given further.

“Big Data is a collection of massive and complex data sets and data volume that include the huge quantities of data, data management capabilities, social media analytics and real-time data” (Ishwarappa & Anuradha 2015, 319.) Big Data analytics involves the study of large amounts of information. Big Data is heterogeneous digital data; data sets are measured in terabytes or petabytes. To process Big Data, innovative methods and techniques are used. These services are developed based on artificial intelligence technology.

Big Data technologies and services are evolving largely due to competition. In this regard, the most successful Big Data projects are implemented in areas of high competition. Using Big Data allows businesses to develop competitive advantages, reduce decision-making time, and make marketing more personalized.
Big Data can be described with the use of five main characteristics (or so-called 5Vs of Big Data): Volume, Velocity, Variety, Veracity, and Value. These features are described as per Ishwarappa & Anuradha (2015, 320-321) in detail below.

- Volume. Analyzing Big Data, the volume factor seems to be the most important. Big Data is, firstly, large volumes of information. The volumes of data used in business and other fields of activity are constantly growing. At the same time, data analysis takes place over the entire volume, not over any single fragment or sample of data.
- Velocity. Big Data is an ever-changing and updated array of information. The rate of change of data is the second important characteristic of Big Data. Static data can be analyzed faster than constantly updated data. Therefore, the speed of data processing is extremely important – it must be fast even in the case of analysis of constantly updated data.
- Variety. The complexity of storing and analyzing Big Data is increasing since data arrays can be heterogeneous – data massive include files of different formats, unstructured documents, messages from social networks and other data. Big Data needs to be structured.
- Veracity. Because of the large amounts of data and the constant changes in Big Data, it is unsurprising and normal that not all data is reliable. This feature of Big Data must be considered when working with technology. It is also worth understanding that the data need not be completely reliable in all cases. In addition, the quality of data in a single dataset can vary significantly.
- Value. Working with Big Data technology and services implies investments in IT infrastructure, in storage and data processing systems. If a company cannot work with Big Data in such a way as to get additional profit from technology, it makes no sense to build the potential of Big Data. Also, this feature of Big Data means the value that technology brings. The potential value of Big Data is huge. At the same time, the data itself is useless until it is used in the most appropriate way in a particular business case.

To sum up, Big Data gives multiple opportunities for comprehensive analysis of large amounts of information. Management of Big Data is carried out according to the following scheme: collecting information, organizing information, analysis, and conclusion. Big Data is used by businesses to improve targeting and marketing in general. 5Vs of Big Data illustrate the ambiguousness of this technology. At the same time, 5Vs are constantly increasing – data volumes now are huge, data are arriving more quickly, and so on. The growth of data's volume, velocity, variety, veracity, and value changes the approaches to marketing information analysis. (Verhoef et. al. 2016, 48).

The data types that are included in Big Data massive are texts, schemes, photo, video, audio, and other information. Big Data is a new and – in some respects – quite radical paradigm of data perception and analysis (at least, Big Data contradicts traditional views on privacy and ethics of communication and data usage). This technology is a global trend that shapes the future. The traditional mode of data has been replaced by a more complete and flexible system – and this global change is perfectly illustrated by 5Vs of Big Data.

3.2.2 Big Data as a part of Data Governance

Both Data Governance and Big Data Governance are united by the feature that the improvement of these management systems takes the business to a new level, improves them. Using outdated methods of Data Governance, companies inevitably fall into a crisis. To understand why a business needs Data Governance and Big Data Governance as its kind, it is needed to trace the history of the company, the evolution path of the business. Without governance, business data is out of sync, with little inner connection. Such data state prevents the company from understanding itself and its development path. Thus, Data Governance and Big Data Governance can help to optimize business processes. (Sarsfield 2018, 17-18).

As Aunimo et. al. (2019, 189) points out, a start-up company can manage Big Data and predictive models based on this data using five key aspects of Big Data Governance: data privacy, security, availability, usability, and integrity. These five fundamental aspects help build a successful Big Data based business. Big Data as part of Data Governance can be used to develop a product and business, to build forecasts and personalize a product to meet the needs of each user. Further, these key aspects according to Aunimo et. al. (2019, 189) will be considered in detail.

- Data privacy. Ensuring the confidentiality of user data is important in all areas of business. This is especially important for IT businesses – if users feel that their privacy has been violated, they behave differently on the Internet and feel anxious. Confidence in business affects the willingness of users to provide their data to companies.
- Security. A security issue is just as important as a data privacy issue. The feeling that the business is safe and protects user privacy affects its commercial success. Companies must decide exactly how they protect or will protect the safety of users.
- Availability. Data accessibility means that authenticated users can access data or software upon request. In other words, the company provides the right people with temporary access to data.
- Usability. Big Data organizations must meet the needs of the business (this is one of the main goals of data management in general). The ease of use of Big Data ensures that it will be monetized.
- Integrity. According to this principle, only authorized people can make changes to the data. The purpose of this principle is to prevent unauthorized use of data. Big Data integrity means that the data used by the business is trustworthy. Because Big Data is less reliable than traditional organization data, there are many potential problems with Big Data integrity.

Big Data Governance framework is especially important when a company just starts the business activity and/or develops data-driven software. According to Aunimo et. al. (2019), Big Data can be managed to optimize business processes both in start-up companies and at the product development stage. A proper Data Governance framework is essential in an authentic business environment. However, this structure cannot improve the business if quality Big Data Governance is not carried out. Without the use of Big Data, data-based processes cannot be beneficial to the business, as they can violate customer confidentiality, have dubious reliability, or may not meet the needs of the business. If a company successfully manages big data, it can become the most valuable asset of the business.

The Data Governance field has emerged in connection with the emergence of Big Data. Big Data differs significantly from the traditional type of data, and it is impossible to manage it without special software or infrastructure. The need for Big Data Governance is currently felt by large and start-up companies. Big Data analytics plays an important role in enhancing business competitiveness and entering the market; by studying data from multiple sources in large quantities, analytics can better understand the business processes and consumers of a product or service than in the case when the data is analysed in a traditional way – for example, when a single data source is analysed. (Aunimo et. al. 2019, 180).

Because Data Governance involves processes that provide formal management of important data assets across an organization, people can trust data when it is governed. However, not all companies use this kind of Data Governance as Big Data Governance.

This is due, firstly, to the fact that the sphere of Big Data Governance is not yet sufficiently developed, and secondly, Big Data cannot be managed with the help of traditional business analytics. Despite this, based on the trend of increasing the role of Big Data, it can be predicted that soon companies will manage Big Data in the same way as they manage all other data. Before establishing a Big Data Governance system, the company should already have Data Governance. (Aunimo et. al. 2019, 181).

Big Data provides opportunities for improving products and services, for planning and implementing innovations. Thanks to the improvement of data collection and analysis systems, businesses can track user experience and the behaviour of potential users, as well as optimize solutions for any innovation. Thus, the analysis of Big Data provides additional advantages compared to data analysis, improves business processes in many areas.

With Data Governance, data becomes a business asset. Since Big Data is not static, it can be considered both as a resource and as a process related to interactions between entities that supply and work with data (El Bassiti 2019, 155.) Big Data is potentially associated with a lot of problems, so effective management is important to optimize Data Governance in general.

Data Governance makes the business more literate and turns data into an important resource for development. A comprehensive Data Governance program includes Big Data management, which is especially important for business success currently. Now the problems of Data Governance and Big Data Governance are problems not only of IT business but of any business that wants to succeed and to get competitive advantages.

3.2.3 Available tools

Big Data services are becoming more popular in recent years because these services can help businesses to understand their customers, to reveal essential information about the target audience. Competing services fight for consumers' attention and constantly improve their functions; assistance of Big Data cloud providers is frequently used for various purposes now. Therefore, a comparative analysis of a few popular Big Data services will help to clarify how the market of such services works.

Google Big Data analytics service, Amazon Web Services, and Microsoft Azure Cloud Services have been chosen for comparative analysis because, according to Gartner, these services are among the best vendors of data science and machine learning platforms. With these services, experts, applications developers and citizen data scientists can analyze their Big Data effectively. If we limit possible Big Data solution vendors to

those, who have cloud infrastructure to natively support data collection from virtual machines, according to Gartner we are limited to these three vendors as they are also the only leaders in this area. In addition, these services enable integration with Microsoft and Linux servers, and their virtual platforms are physically located in the EU, which is very important in current state of data privacy laws and regulations.

Google as a searching system itself is a training machine for artificial intelligence because users by their search queries make the system more perfect, and Big Data technologies are integrated into the most popular searching system in the world. Also, Google uses Big Data to improve the world – for instance, to protect the environment (Wang, 2016.) Nevertheless, Google as a huge data massive is not a subject of our analysis. Below, Google Big Data services created for customers (individuals and companies) will be regarded.

Google Big Data analytics service, or Google Cloud Platform (GCP) offers to its users "a comprehensive and serverless data analytics and machine-learning platform" (Big data analytics 2019.) GCP has a wide range of Big Data subservices (at least 11 – among them are BigQuery, Cloud Dataflow, Cloud Dataproc, Cloud Pub/Sub, Cloud Data Fusion, Cloud Composer, Data Catalog, etc. (Big data products 2019)). So, advantages and disadvantages (pros and cons) of the entire platform, not separate subservices, will be summarized.

Pros (Big data analytics 2019):

- Google service overcomes the main limitations of Big Data services: scales, performance, and cost efficiency. A secure and automatic platform that works without servers.
- Simple and clear interfaces. The platform focuses on analytics, not data organization. Because of the convenience of the service, users can exceptionally work with data and forget about managing servers.
- Easy access for customers and high security of services. REST-based APIs for easier integration with other apps.
- Extremely high speed of work and development. Gigabytes and petabytes of data may be analyzed thanks to BigQuery (fast speed of ANSI SQL).
- Prices are friendly to users – customers pay only for the functions and resources they use. Also, there is a discount system on the platform. Service has some of the lowest prices on the market.
- Platform suits for large companies. Among customers of platform are popular e-commerce brands (such as Snapchat).
- Google tutorials help to understand the platform quicker.

Cons (Gandham 2019):

- Some services are badly managed, have outdated and limited functions.
- Core GCP products (for example, BigQuery, Datastore, and Spanner) have limited observability and customization. If a business uses its own ways of managing working processes, these products can work imperfectly.

- Many GCP services are underdeveloped, they work in a beta stage for years. SDKs are broken and documentation is poor.
- A few Google services are “global”, which means that they are created for working in various regions. However, they have not enough resources to work properly in all cases.
- Imperfect system of customer support and sales.

Thus, storage services available at GCP, or Google Cloud analytics products, offer multiple opportunities to its users but still have limitations and imperfections. The most notable feature of GCP is its base – the service is created on the ready foundation: Google’s artificial intelligence. Users of GCP get access to the industry-leading technologies and services created by Google with the use of data collected for the 20 years of the brand’s history. The infrastructure of GCP subservices is reliable, secure and scalable. GCP also offers a serverless data platform to its users. Among key GCP products are business intelligence, data lake, data integration, stream analytics, etc. (Big Data Analytics 2019.) Despite all the imperfections of the platform, the key role of GCP on Big Data services market is undeniable, since Google is the world leading brand associated with artificial intelligence and Big Data.

Big Data can also be analyzed via Amazon Web Services (AWS). This is the second service for comparative analysis.

Pros (Data Lakes and Analytics on AWS 2019):

- Global reach and scalability.
- All Amazon services have API which helps to create immutable IT infrastructure for customers.
- Users may focus on creating codes while the other aspects may be neglected.
- More productive and appropriate for startups. Easy to use even for unexperienced customers.
- There are many tools and manuals to make work with AWS more convenient.

Cons (Pros and Cons of using Amazon Cloud Services 2018):

- Security limitations.
- The prices are higher than on GCP and Azure.
- No tutorials. Services are quite difficult to understanding without additional information. Configuration of services according to company’s needs may require the help of partners.
- APIs are multiple but far from perfection.
- Additional fees for customer support.

So, AWS is a cloud-based platform created for businesses that unite inter-connected services. Some of AWS services help companies to develop their own cloud-based solutions.

The third service that will be analyzed in this subchapter is Microsoft Azure Cloud Services. Azure is a cloud computing service created for programming, testing and deploying apps. So, software, platform, and infrastructure capabilities are united in this service.

Pros (Top 7 Advantages of Using Azure Cloud Services 2018):

- Available globally. The speed is higher than on the most of similar services of competitive brands.
- High security standards. Security is a primary built-in feature of Azure
- Azure tries to reduce the risk of data losses maximally.
- The flexibility of scaling. This is one of the main features of Azure.
- Any framework, language or tool can be used.
- Business can build well-developed, hybrid infrastructure with the use of Azure.
- Multiple artificial intelligence services.
- Recovery in the case of disaster.
- Appropriate for startups.
- Cost-effective service.

Cons (Gaille 2018):

- Constant management of service is required to make an effect of its usage. The service does not help users to manage their data.
- Platform expertise is needed to use Azure.
- A single vendor strategy is proposed by the service. Despite its convenience, such a strategy increases risk.
- The speed of work may be lower in some regions (nevertheless, as it was mentioned above, the average global speed is extremely high).
- Some businesses may have problems with the ease of access.

Thus, Microsoft Azure, like GCP and AWS, sets various global services to businesses to reduce IT costs, make digital transformation more rapid and increase the flexibility of data. GCP is better for large companies with organized business processes. AWS and Azure are more suitable for startups because of their operational simplicity. Microsoft Azure is chosen for further empirical research. Azure was chosen because of good documentation and because the company described in the empirical study uses many Azure services. For all three systems reviewed, various tools and services work best, but Azure turned out to be the most convenient for the selected company.

3.3 Importance of pilot project – proof-of-concept

Any business project involves changes in mental and/or physical environment. Regardless of which new product or business appears on the market, their launch is associated with potential risk. To reduce this risk, pilot projects are being implemented. Pilot projects are also important when a company tries out new technology. (Zbrodoff 2012).

A pilot project is a trial launch of the product, a small version of a larger one; it is “an initial small-scale implementation that is used to prove the viability of a project idea” (What Is the Difference between a Trial and a Pilot? 2019.) Thanks to the pilot project, the company can understand that the product in preparation for release stage needs to be changed. The concept of a pilot project is also used in the field of IT. A software pilot project is a test of new software, its local verification. For example, new organization software can only be used in one part of the organization, and not across the entire company. A new software product can be offered to users in a demo version, etc.

In addition to the pilot project, a pilot research (or a pilot study) is valuable for future product. The concept of the pilot research is used in the humanities. The pilot study is the study in which a specific research tool is tested: “The pilot study will confirm viability and scalability and enable proposed processes and procedures to be tested... It also enables the benefits to be tested and a more reliable investment appraisal to be created for the main project” (What Is the Difference between a Trial and a Pilot? 2019.)

The pilot research can also mean a small-scale launch of a product, its trial version, made in preparation for a full study. An experimental study can show the weaknesses of a project, try out a separate research tool in practice, avoid project risks or minimize them (Roland van Teijlingen & Hundley 2002, 33.) Like pilot projects, pilot studies are funded (Turner 2002, 4.) We will use the concept of the pilot research in the meaning of theoretical analysis before implementing the pilot project.

Pilot projects are often used to test new technology, software, or another product. They allow businesses to see the technology in action and improve it for getting a successful result. In addition, pilot projects help employees to become familiar with technologies and influence decision-making regarding the final type of product, as well as how to implement the product. The disadvantage of pilot projects is that if the audience’s attention is not directed to them, the projects may show false inefficiency, or, if the attention is drawn, the product that could become successful will not be implemented or will take a different look. Thus, a pilot project can be an obstacle to change. In addition, excessive attention to the pilot project may create the wrong focus for the whole team.

Despite some shortcomings, pilot projects can be an effective means of monitoring the main project. Thanks to the pilot project, the business can understand what is working well in the product and what needs to be changed. Pilot projects are becoming a source of valuable information about product capabilities. In the IT field, pilot projects help to under-

stand whether the release of software in the form in which it was designed is the right decision. IT products are tested in terms of their functionality and relevance to the needs of the audience.

Software developers or third-party vendors can participate in the pilot software project. When starting a pilot project, IT specialists are engaged in its network administration, and the training and support group collects data on user problems. A group of users is involved in testing an IT product.

Summarizing all previous and adding key points by Gupta (2018) and English (2019) to ensure the success of the pilot project, it is important to imply the following rules for its organization:

- Work mainly with the most important problems for the product.
- Make sure that the pilot project has enough degree of support. It is important not to go to extremes – a team can learn from the mistakes of the pilot project.
- The purpose of each pilot project should be clear. Pilot project support groups should be aware of their responsibilities. If the goal was not achieved during the implementation of the pilot project in the IT sector, this may mean that the project uses the wrong tools or that the expectations from these tools were too optimistic.
- It is important to know what indicators will be checked in assessing the effectiveness of the pilot project.
- Track all the problems that arise during the implementation phase of the pilot project. It is also worth paying attention to those aspects that are implemented without any problems. This may be an example for other areas.
- Often visit the pilot site, communicate with product users about what problems they are facing, what changes they want.
- People working on a pilot project should analyze failures and their causes. Failures should not be repeated.
- It is important that people working on a project discuss with each other what does not work in the project. Project management should carefully assess these issues without blaming specific people.
- Encourage people to use new technologies.
- Make system changes to work on pilot projects so that the next project does not have the same problems.

Each pilot project is a unique experiment. Using a pilot project, company management finds out what changes should be made to the product which production is being prepared. Before starting the pilot project, it is necessary to determine its desired result, set the scope of the project software and request information on possible solutions. A customized project monitoring system will help identify all the problems of the project and find ways to solve them. After the time allotted for the pilot project, the team needs to evaluate how the pilot launch went and write a report.

Thus, pilot projects allow companies to minimize risks and waste of time and money when starting a project. In the IT industry, pilot launches are testing new programs or technologies. Among other advantages of this method, it is important that pilot projects help evaluate options.

4 Preparing for Proof-of-Concept

In this chapter are described all preparation phases before starting actual pilot projects. It starts with description of target organization, it's current state and customer groups. After that the chapter goes through forming organization's vision about projects importance and desired state of data governance program, stakeholder selecting, discovering available data resources, forming data governance team, defining processes for the projects and initial plan of implementation.

4.1 Target organization

FIMX Oy is innovation leader in digital property information services. It provides sophisticated tools for operating with property information using SaaS (Software as a Service) approach available thru any web browser. For FIMX Oy being in this role has been the everyday now for a decade.

In their service they have currently about 6000 organizations and their users make use of FIMX service in different roles. FIMX customers are large property owners, condominiums, real estates, managers as well as different service providers (maintenance, cleaning, hvac, electricity etc.). The electronic service- and maintenance request channel (JULMO) has brought the real time resource planning also for the reach of hundreds of thousands of property users, tenants and residents.

Over eight million service requests have been handled in FIMX service. The complete process from announcer's public form and different phases of the work to accepting the actions, invoicing and customer feedback from the announcer are in the center of the service. All the different parties handling and following the process see and communicate the same phases of the process and all of them have always 100% accurate information.

FIMX maintenance- and service request process creates customer satisfaction key indicators and response for thousands of properties. All together the smileys for the work, its quality and efficiency have been clicked for over 420 000 times! The evaluations are immediate feedback from the announcer and subscriber for the quality of the actions. Most of then – especially the negative ones – include also a feedback and often a development proposal to improve quality. These have been given in FIMX-service for over 125 000 times.

Features of FIMX service include:

- management of maintenance- and service requests, real time follow up of the work and instant response
- preliminary maintenance tasks (maintenance manual)
- managing archive and attachments, folders
- long term maintenance plan
- consumption reports for electricity, water, heating, cooling…
- device register
- owner reporting, e.g. service production assessment, response times, long term maintenance plan, consumption comparisons, waste reporting
- mobile compatible interface
- continuous assessment of service satisfaction
- electronic channel for maintenance- and service requests (JULMO)
- rescue plan based on the real time data in our service
- map interface for reporting
- QR-codes for service requests, meters, devices, spaces etc.

4.2 Agreement with organization's management

Before starting actual project planning it was very important to gain common understanding about data governance value with organization's owner and other high-level managers. Without support from the upper management it would be close to impossible to gain resources required for good quality implementation of pilot data governance project. Understanding importance of this step led me to conclusion that it also requires good preparation.

Target organization at the time of writing this thesis consisted of ten people in total. This restriction has its pros and cons. Main disadvantage was in limited possibly human resources allocated for pilot project. Main benefit was very low-level hierarchy, that allowed discussion with anyone in organization in person. Also, negligible hierarchy and small size of organization meant that gaining "green light" for pilot project from the upper management would help a lot in allocating human resources for it.

As a team leader of one of IT-teams within the target organization I have pretty good understanding about all current projects and daily small talks with development manager. These factors made easy to start conversation with development manager about current state of organization's data governance and possibilities to extract new value from it. Initial conversation about the subject was kept and as a result of that conversation several conclusions were made:

- We have a lot of data and many different types of it
- The data we have can be made more valuable improving data governance approach
- Possible pilot project's target customer groups can be building managers and building owners as they are assumed to be most interested in easy-to-read insights of data
- Insights from processed data can support software development

It was great win to gain support from development manager as he has a lot of influence on planning big projects in target organization. The next conversation took place at weekly meeting of organization's management and permission to start pilot project were granted. In addition to that, other possibilities for making our in-house data more valuable were discussed. This discussion led us to agreement, that we use our data for improving organization's activities. From management point of view, possible next projects can use data insights to support billing activities, better understanding on our services usage and user profiling as a support of overall planning and software development prioritizing.

These two conversations ensured that organization's management promised support to this project and were curious about its results. We got permission to further project development in cooperation with development manager and possibility to use other organization's human resources. Also, we promised to report weekly about current state of the project.

4.3 Discovering available resources

To start planning data architecture and data models of new project it was important to discover overall organization's data architecture and available data models. This step was obligatory as no single project can be totally isolated from other projects. Especially this is true for data governance projects that use existing data and produce new knowledge from it.

This step required awareness of current state of data storage and operations. For this purpose, existing technical architecture model was re-evaluated and blueprinted as shown in Appendix 1 (confidential), Figure 1. This model helped to list data resources that exists within the architecture and so are processable using existing tools. In addition to that several possible data resources where defined as potentially useful but out of scope of technical architecture, so they require implementing new tools to become processable. Both data resource lists are shown in Appendix 1 (confidential).

Discovering current state of data warehousing and business intelligent activities was logical next step to form insight on tools that are already in use in target organization. Existing technical architecture also provided better understanding of technical tools that were used within the organization. This knowledge helped to list tools that can be used in this pilot project. For better understanding possibilities and limitations of these tools organization's inhouse technical expert were interviewed. This interview resulted in forming list of available technical tools and their specifications. The list is also shown in Appendix 1 (confidential).

4.4 Planning tools and investments

Before starting actual project, we needed to do a commitment about possible tools used in it and possible investments the project may require. The project was implemented in iterative way, which meant that all tools and investments couldn't be specified very detailed, but we still needed to form our bounds and possibilities. These bounds could then be approved by organization's management and act like a guideline for actual project implementation.

As mentioned previously there is multiple tools for building Big Data solutions as an in-house environment or using cloud environments. Building in-house Big Data environment have some benefits, but in our case, it was potentially costly and very complicated as we had lack of experience and furthermore limited human resources allocated for the pilot project. Using cloud solution from one of three main providers on the market Amazon, Google or Microsoft was seen the most cost-efficient way. Comparing these three cloud providers and their solutions guided us to conclusion that there are just minor differences between them and close to no differences when trying to compare solutions from Amazon and Microsoft. Target organization already had experience in using various services from Microsoft and an active Microsoft Azure subscription with multiple use cases. In comparison to that, there was no experience in using services from other Big Data solutions provider. Because Microsoft Azure was one of the leaders in the market and we were already using some services at that platform, it was logical choice as a platform for pilot project.

Because Microsoft Azure uses pay-per-use subscription plan with many variables, it was hard to analyze total costs for pilot project without detailed plan. Iterative implementation added more complexity in total cost calculations. In comparison to that between iterations of project it was easier to plan financial and human investments for next iteration. Suggestion for using this investment model was approved by target organization's management so we were ready to start actual project detailed planning.

5 Implementation of Proof-of-Concept

This chapter describes all steps of implementation of pilot projects within organization's Data Governance program and their current status at the moment of writing this thesis.

5.1 First pilot project: visualization of multidimensional data

Target organization generates a lot of data about buildings and their usage. Based on that data target organization implement various reports for building managers and building owners. Most of these reports provide visualization and detailed report about some specific data collected from buildings usage. These reports can be grouped by single building and by user-managed building groups bringing comprehensive information for various user groups.

Most of the reports are displaying data about some particular data and there was no previous single report for visualizing main information from all of them. There was a lot of conversation from time to time about benefits of report where our clients can see at a glance holistic information without digging into specific reports. Initial interview with development manager about possible pilot project for valuable Big Data solution led us to decision that this kind of report can be a good candidate, because it has all main aspects for that:

- It can be very valuable for our clients
- It requires a lot of various data processing
- It requires excellent data visualization to be useful

Based on these interview results this kind of report was selected as first pilot project and more detailed project planning was started.

5.1.1 Planning the project

Actual project planning started from taking interviews within target organization and collecting all desired features of this report solution. Interviews were in form of small talks and conversations in organization's chat. Interview notes were collected as list of possible features and then validated with all participants to ensure that nothing important left unattended. In addition to that, input data about possible features was collected from email conversations with organizations clients. Final list of all desired features is shown in Appendix 2 (confidential). As can be seen from this list, desired features were not only visual reports but much more, including multiple automated tasks, notifications, lists and alerts. However, it was agreed that in first iteration only visual reports will be released and scope of first iteration should be limited only to them.

During these interviews raised understanding that final product of this project must be a part of our solution for our clients. That understanding launched search of data visualization tool which can be seamlessly integrated in our solution. Quick research about available visualization tools showed, that although there is a lot of tools in the market, filtering them by parameters like free for corporate usage, variation of charts and graphs, good documentation, good customization and easy to implement narrowed selection to just couple of them. Current market leader seems to be D3js and other frameworks based on it with its huge community and sophisticated customization possibilities, although learning curve can be somewhat difficult. Two another possible visualization tools were Google Charts and ChartJs. We had previous experience using Google Charts and in some special cases its customization limits were reached. Also, for this project was decided to test some new visualization tool, so we quickly end up testing possibilities of ChartJs in our pilot project.

Limiting scope of desired features to visual reports only simplified implementation very much. All data, that we needed for this project can be found from organization's own structured and well documented data resources. Result of this project planned to be part of existing web application. These two boundaries made also data privacy and data storage and operations straightforward because of multiple reasons: there is no need to transfer data to external platforms, existing data privacy tools and settings can be reused, processed data would not generate any new person related data.

Next step was planning data model of pilot project. Data model of final report was decided to be same for all transformed data to simplify adding new data resources even after first implementation. Data model of transformed data planned to have relational structure and to have following attributes:

- Unique building identifier
- Timespan of collected data
- Collected data type
- Calculated value
- Quantity of data rows used for data calculation

All calculated values of data model should be also unique to ensure data quality. Therefore, limiting uniqueness of calculated data should be based on combination of building identifier, timespan and data type. As building identifier is already unique and timespan can be ensured during extracting phase of ETL process, there was only one new attribute, that must be declared and specified in metadata of data model – collected data type. Collected data types was listed as enumeration. This enumeration combined with data model

and descriptions of source data formed a good base for metadata. Final form of metadata was left to be written out after first implementation of the project.

After specifying data model, it was data integration and interoperability turn. For this step I created initial ETL-model, which is shown below. For organization's data privacy reasons actual collected data types and report names are replaced with numbers.

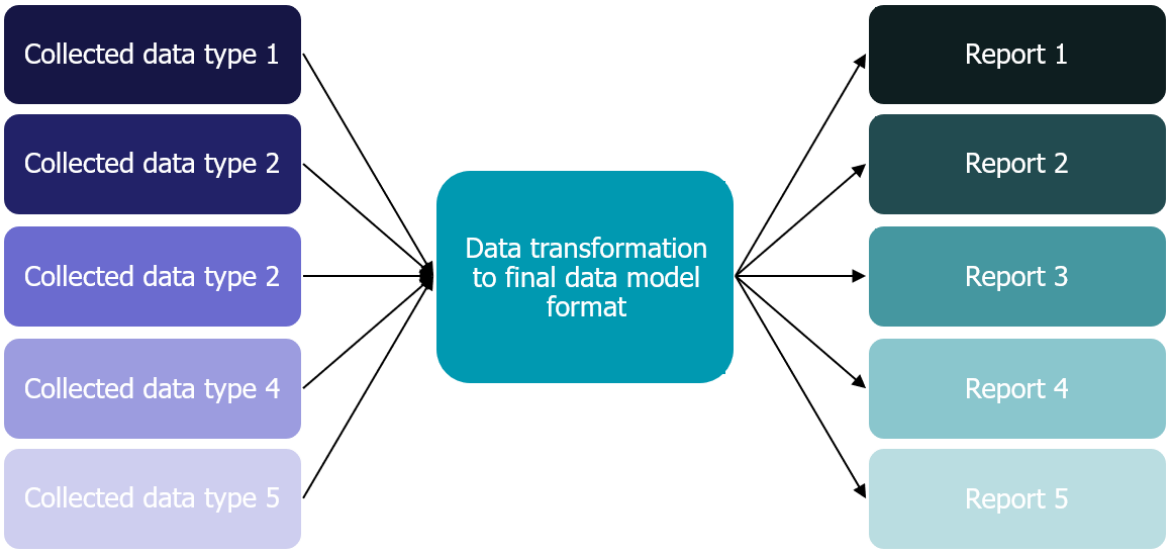


Figure 3. Initial ETL model of first pilot project

This ETL-model combined with data model made clear, that collected data especially unique building number cab be considered as master data, when transformed data and reports based on it formed reference data. This also formulated statement, that data quality of this project is based on two main aspects: quality of collected data and quality of data transformation.

Deeper research on source data quality and precise rules of data transformation were left on implementation phase of the project. For that reason, data architecture of this project was also left to be written out after first implementation of the project, although relation between the project and other organization's solutions and EDM was confirmed.

Putted altogether selected tools, note about data privacy, note about data storage & operations, data model, initial metadata, data integration & interoperability, data quality rules and initial data architecture formed guidance for implementation and actual project implementation started right after it.

5.1.2 Implementing the project

Actual implementation started with discovering available data sources and analyzing their quality. One of aspects of final output was to provide information about wrongly inputted or missing data, so at this point only technical data quality issues could be critical. Comparing database model and actual current database state showed, that database model is very well formed and there are no technical issues affecting data quality. However initial data processing attempts showed that there are big differences in data quantity from building to building. This finding strengthened understanding, that final reports based on processed data should highlight these data quantity anomalies to be more informative. Another conclusion derived from initial data processing was that this step can be done using existing tools within existing environment. Also, this step showed that fine tuning of data processing can be done separately from fine tuning of final reports and their user experience, even at some point it is very fruitful to validate processed data against visualization. It allowed to divide implementation to two separate tasks: data transformation implementation and report visualization implementation. These tasks were allocated to two project participants to speedup project implementation.

When data transformation implementation was ready, it consisted of multiple similar tasks. Each task was responsible for collecting and transforming data into format suitable for further processing during visualization part. Putted in a batch job all tasks took eight to nine minutes in total to process data for past three years. It was possible to reduce processing time collecting only data for time period that was not processed previously, but one finding restricted this possibility. Comparing results of processed data showed, that new input data can affect some data in the past, the root cause was found, and it was acceptable and easily explainable. Further data validation could be made only after implementation of visualization.

Basing on previous conversations with customers and other reports usage statistics it was decided that the visualization should have three levels of timespan selecting: year, quartal and month. Although minimal level allowed to schedule data processing only once a month, for more accurate and up to date data it was reasonable to process data every day. As data processing batch job took 8-9 minutes and was quite resource costly, it was planned to run it at night, when other resource usage is at it’s minimum. Daily processing made possible one of future features – fast customer notifications about rate changes in reports which they mark as important.

Visualization implementation using ChatJs was quite easy because of good documentation and very simple framework syntax. For faster response from development team and inhouse stakeholders, it was decided to implement simple report based on total counts from processed data as the most straightforward data-to-visualization implementation. Alpha version of this report is shown in Figure 4. It had no navigation, no possibilities to change timespan and even no localizations, but it showed easiness of building visualizations from processed data. In addition to that it showed very good speed of retrieving data and building report after data is retrieved.

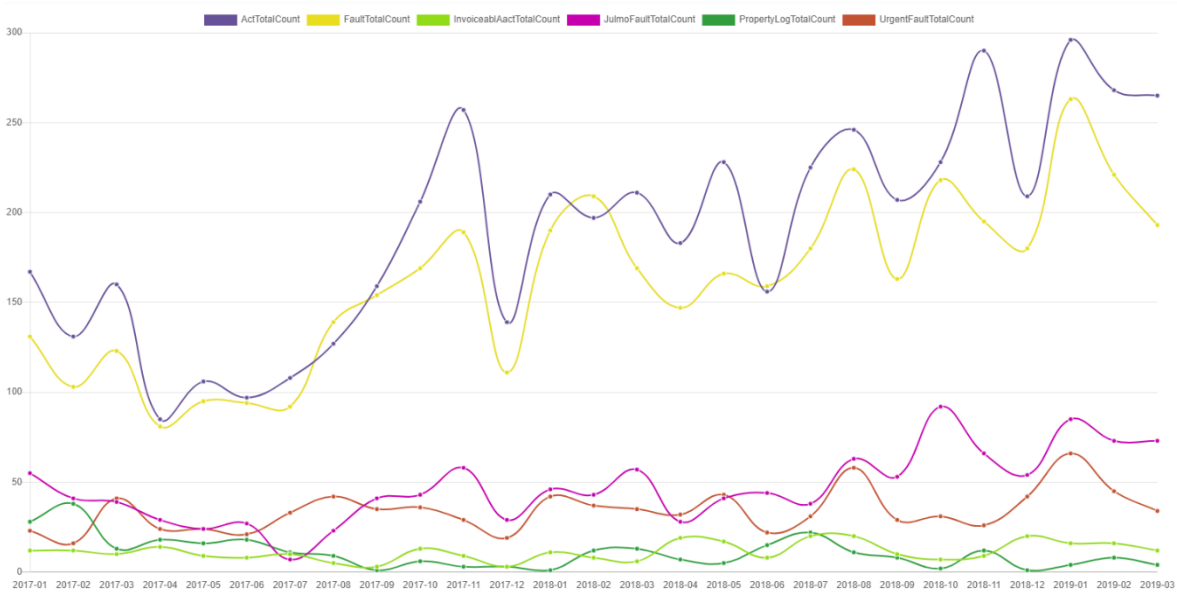


Figure 4. Alpha version of visualization of processed data

During the next meeting of inhouse stakeholders some of user interface features of final report was planned. First it should have easy navigation between reports, between building portfolios and between buildings within portfolio. As there will be multiple dataset per report, it should have feature to hide less important datasets and pre-set option to highlight more important datasets. There must be two graphs at building portfolio level to show summarized results of portfolio and building comparison within portfolio. In addition to that building comparison can be shown in form of table for those, who are interested in more detailed data. This table can be also used as navigation from portfolio level to single building level. As some of our customers has building portfolios as big as hundreds of buildings per portfolio, it was valuable to add possibility to filter report buildings by city, region, address and even by real estate service provider company.

Implementation of user interface features was made simultaneously with improving data processing tasks. As visual part started to be more mature it brought up bottlenecks in processed data. Each improvement of processing tasks was validated and approved by all

members of development team and tested on various building sets. This technique helped avoiding human errors. Each visual effect and feature were also implemented using same technique and in addition to that there was two demonstrations of beta versions for organization's management. These demonstrations received very positive feedbacks and ensured that development team and organization's management shared same understanding about desired outcome. Additionally, one new feature was introduced. Because graphs generation from processed data was very resource effective and fast, it was chosen to bring some of them directly to the building portfolio and building main pages to show key trends to customers as soon as they start to investigate any available report.

When all localizations, planned user interface features, reports defaults and data processing optimizations were ready, project development team published final version of data model. Demonstrated to organization's management project was approved to be published to customers. Further project development was planned after collection of customer usage statistics and customer feedbacks. Preparation to actual publishing included composing in-app notification and public articles on organization's main website. Links to these articles can be found in Appendix 4.

5.1.3 Short reflection of current state and future possibilities

Sample of user interface at the moment of writing this thesis is shown in Figure 5. It contains all features planned during first iteration and gained very positive feedback from customers. Organization's management was excited about project development team outcome and this pilot project was accepted as successful.

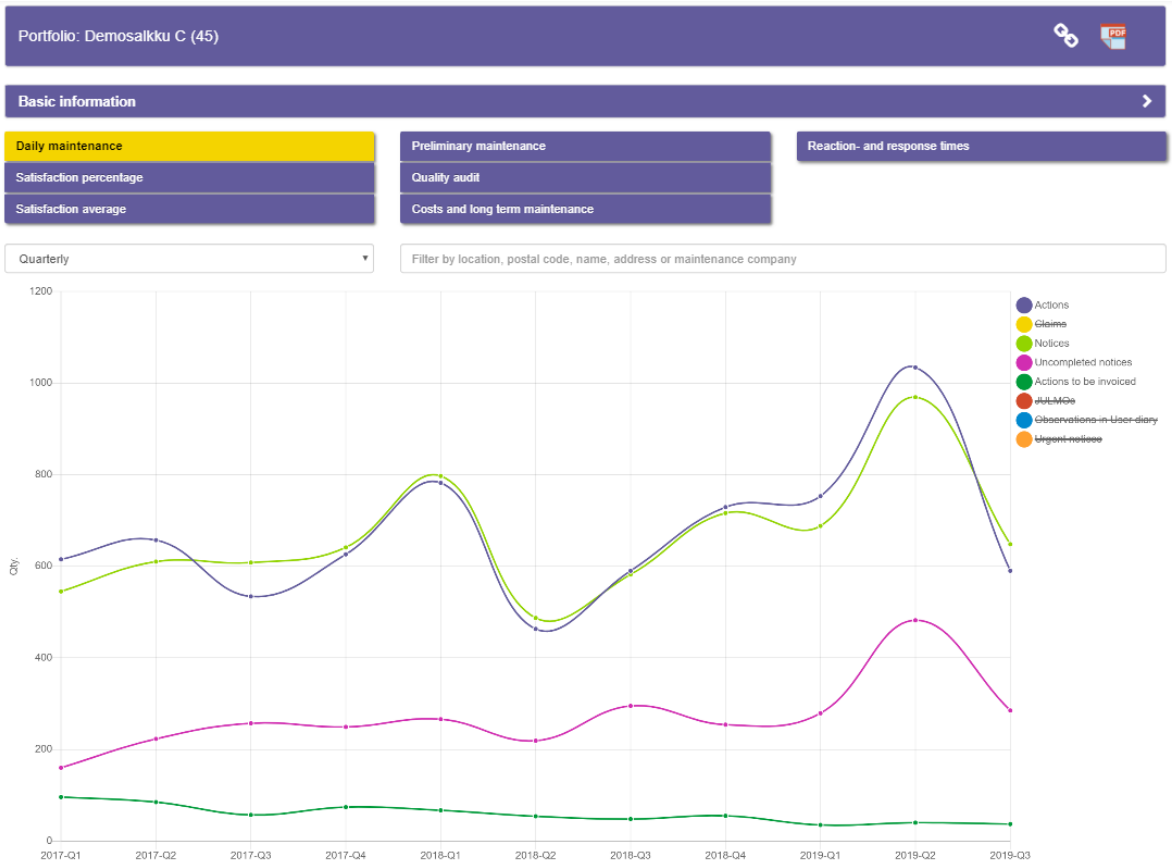


Figure 5. Final version of user interface

Customer usage statistic collection was still in process. It was done partially manually and took too much time from combining usage data to report that can be shown to organization's management. This bottleneck was acknowledged, and organization's management granted permission to start new project by same development team to fix this issue. More detailed description of this project is shown in next chapter of this thesis.

5.2 Second pilot project: extracting valuable data from software logs

Target organization delivers value to its customers using SaaS (Software as a Service) approach via web applications making it available from any device with Internet connection and any modern web browser. At the moment of writing this thesis there was two main web applications available to all customers: one hosted at m.fimx.fi domain and another hosted at pro.fimx.fi domain. These web applications produce usage logs at various levels and in various formats. First level of usage logs is produced by web servers of these applications. Second level of usage logs is produced by web applications themselves using in-application custom logs. Third level of usage logs is related to database objects and their history. Every level of produced logs is using its own format and all of them are stored separately making combining them a big challenge.

Validating results of previous sub-project using software logs required platform, which can combine all levels and formats of logs into single real-time or near-real-time report relevant for organization's management. Short conversation with organization's CTO and CEO resulted in common understanding that developing of this kind of platform would not only benefit this sub-project but can be used for deep comprehensive analysis of both web applications. Furthermore, such analysis can in future support data driven development bringing better understanding about current clients' needs. Target organization's CEO was very enthusiastic about this vision and not only granted permission for starting pilot project but also suggested desired output for it: collecting data about resources that are not used by clients.

This idea gained support from CTO, who pointed out that such knowledge can promote multiple development targets:

- Focusing development on features that are most used by clients
- Reducing codebase by removing features that are not in use
- Improving availability of features which are potentially useful for clients

Support of CEO and CTO ensured that implementation of pilot project can be started.

5.2.1 Planning the project

Actual planning of the project started from discovering data sources, which can be useful from perspective of desired outcome. This analysis showed that three levels of logs from two web applications using different technologies result in six real-time logs and two codebases each having its own format. Variety, volume and velocity of possible data sources was challenging for this Big Data pilot project. Acknowledging this I suggested limiting scope of pilot project only to one of web application, which halved data sources and simplified first iteration of implementation. Organization's CEO accepted my suggestion for proof-of-concept.

Selection which of web applications will be target for proof-of-concept was made based on easiness of implementation. Because one of them was already collecting web server logs and in-application logs which will be used as master data to Microsoft Azure it was obvious to select this application to be first as this can simplify data collection and data processing significantly. The collected reflected only resources that were in use so to provide desired output it needed to be linked to list of all available resources within application. This list of all available resources can be used as reference data. Collection reference data and methods of bringing it to Microsoft Azure portal was left on implementation phase of this project.

When master data and reference data sources were selected, it was time to ensure Data Security and Data Privacy rules are fulfilled. Together with Data Protection Officer (DPO) we verified, that already collected data fulfil organization's Data Security and Data Privacy policies and government's obligations. Microsoft Azure portal allows client in easy steps to determine, in which physical location of their cloud data is stored and who have access to it. At the moment of writing this thesis all data was physically stored in EU and only authorized organization's personnel had access to it. Reference data do not need to have any personal information, but it's cloud's physical location also was planned to be in EU because it simplifies Data Storage and Operations handling.

The next step was to plan Data Architecture keeping in mind Data Integration and Interoperability. For this project possibility to integrate various data sources between each other and possibility to transform data efficiently was a key to success so DII was tightly coupled with Data Architecture. For achieving this goal and splitting implementation into smaller steps Extract-Transform-Load model was made. It is shown in Figure 6 below.

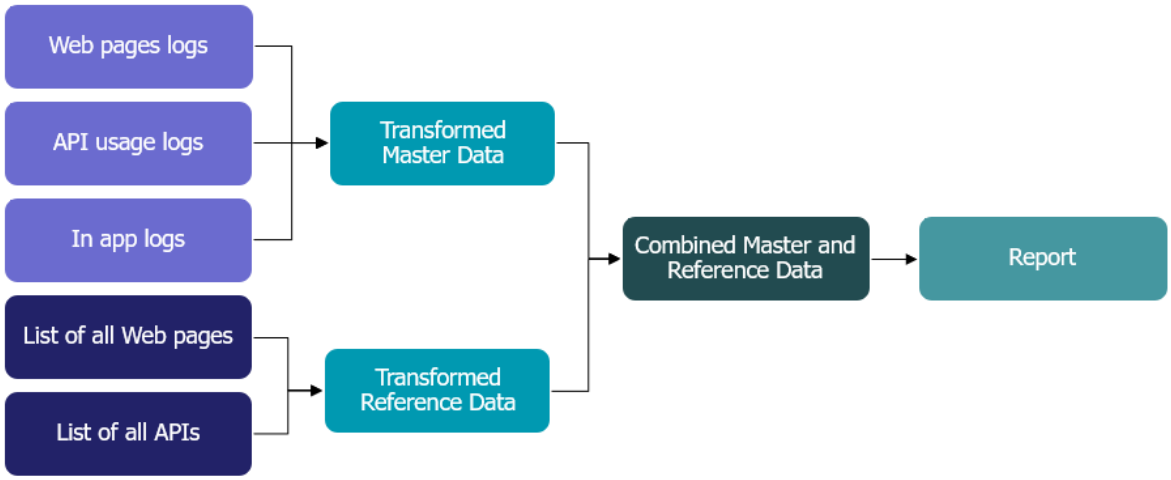


Figure 6. ETL model of second sub project.

All steps of this model were planned to be automated. This Data Warehousing and Business Intelligence challenge required a tool or set of tools to accomplish it. ETL presented if Figure 6. helped to select these tools and made actual implementation straightforward. As mentioned previously, all three log sources were already collected into Microsoft Azure Portal. Inside Microsoft Azure Portal Azure Data Explorer uses Kusto Query Language (KQL) to query structured and semi-structured data. It allows to query any kind of logs and render graphic representation of results. KQL is powerful but also the only query language in Azure Portal that can be used to extract information from collected logs. Therefore, also reference data needs to be processed in form that can be queried using KQL. Fortunately,

Azure Log Analytics supports custom logs collection from various formats and make them available for KQL.

In addition to data querying challenges, automation of process required scheduled trigger and automated reporting. Microsoft Azure Logic App is a set of tools, which support multiple inputs, outputs, logical operations, data transformation, KQL querying, Azure Functions calling and so on. Azure Logic App builder is a visual tool which can be used from Azure Portal simply by drag-and-drop technique allowing to build sophisticated apps without need to write machine code. All these possibilities made selecting Azure Logic App as technical framework for all data transformation, combining and reporting a coherent decision. Planning actual parts of this Logic App was left on implementation phase.

5.2.2 Implementing the project

Implementation of this sub project started with extracting Reference Data mentioned in Figure 6. As list of web pages and API's are not static and change during development of web application, using Azure Log Analytics custom logs for extracted data was rational choice. It allows collecting same types of logs automatically from predefined location. I generated script which checks the code of web application and exports all web pages routes and all API routes to CSV files. These CSV files are then picked up by Azure Monitor which brings them to Azure Log Analytics where data can be transformed for further processing.

As was mentioned previously, Master Data such as Web pages logs, API usage logs and in-app logs were already collected to Azure Log Analytics. So next step was to start transforming both Master and Reference Data to make them ready to be processed for final output of this project. In cooperation with Senior Specialist we formed multiple pre-defined KQL-queries from source data, which were planned to be automatically processed. The queries were validated several times to ensure that they produce correct information processable by next step. After validation we were ready to create information processing app to gain actual knowledge.

Gaining knowledge in a real or near-real time and acting based on that knowledge in a form of sending results to organizations management was a multidimensional task. Developing the tool or set of tools for this kind of task by programming them using organization's own resources could be a very timely and costly process. Fortunately, among big list of different tools in Azure Marketplace, which is available through Azure Portal, there is a tool called Logic App, which is created by Microsoft. It allows to connect many sorts of

predefined input and output APIs such as KQL querying, Azure Functions calls, database operations, email sending, natural language processing and so on. There is even possibility to write your own APIs if comprehensive list of available APIs is not enough. In addition to that it can be triggered by event or schedule, APIs can be chained one after another and there is good set of logical operations, which can be used for acting based on output from any of these APIs. There is also quite user-friendly browser-based tool, which allows to create sophisticated Logical App just by putting API “blocks” together. There was no previous experience in target organization, but it sounded as a good choice for this project.

Creating of Logical App started from forming KQL-query to be triggered by Logic App schedule. Although it wasn’t very complex task, it showed very shortly, that after querying, data should be processed and validated by something, what can be tested to produce correct information from this data. That was the job for Azure Functions. Azure Functions also can be triggered by schedule or event and send output to any API, just like Logic App. But unlike Logic App, they must be written in some programming language. Using programming language for complex data processing has big advantage – data processing can be tested and validated by unit tests and test data. At the moment of writing this thesis Microsoft Azure Functions support C#, JS, F#, Java, PowerShell, Python. Because target organization mainly uses C# in daily work, it was logical to select it as programming language for Azure Functions also as it allows to use in-house resources for writing necessary Azure Functions.

When Azure Functions were ready and validated to work as expected it was time to add Logic App connector for saving processed data. There are many options for storing data in Azure and sending it to any storage outside Azure, or even sending graphical report directly to email. Because results from this data processing Logic App could possibly serve multiple targets, it was decided to save also processed data in Azure Log Analytics and create separate report sending Logic Apps based on demand.

Figure 7 shows final version of Logic App. It starts with scheduled trigger, runs KQL-query, prepares data for first Azure Function, processes data in Azure Functions, prepares processed data for storing and stores it in Azure Log Analytics.

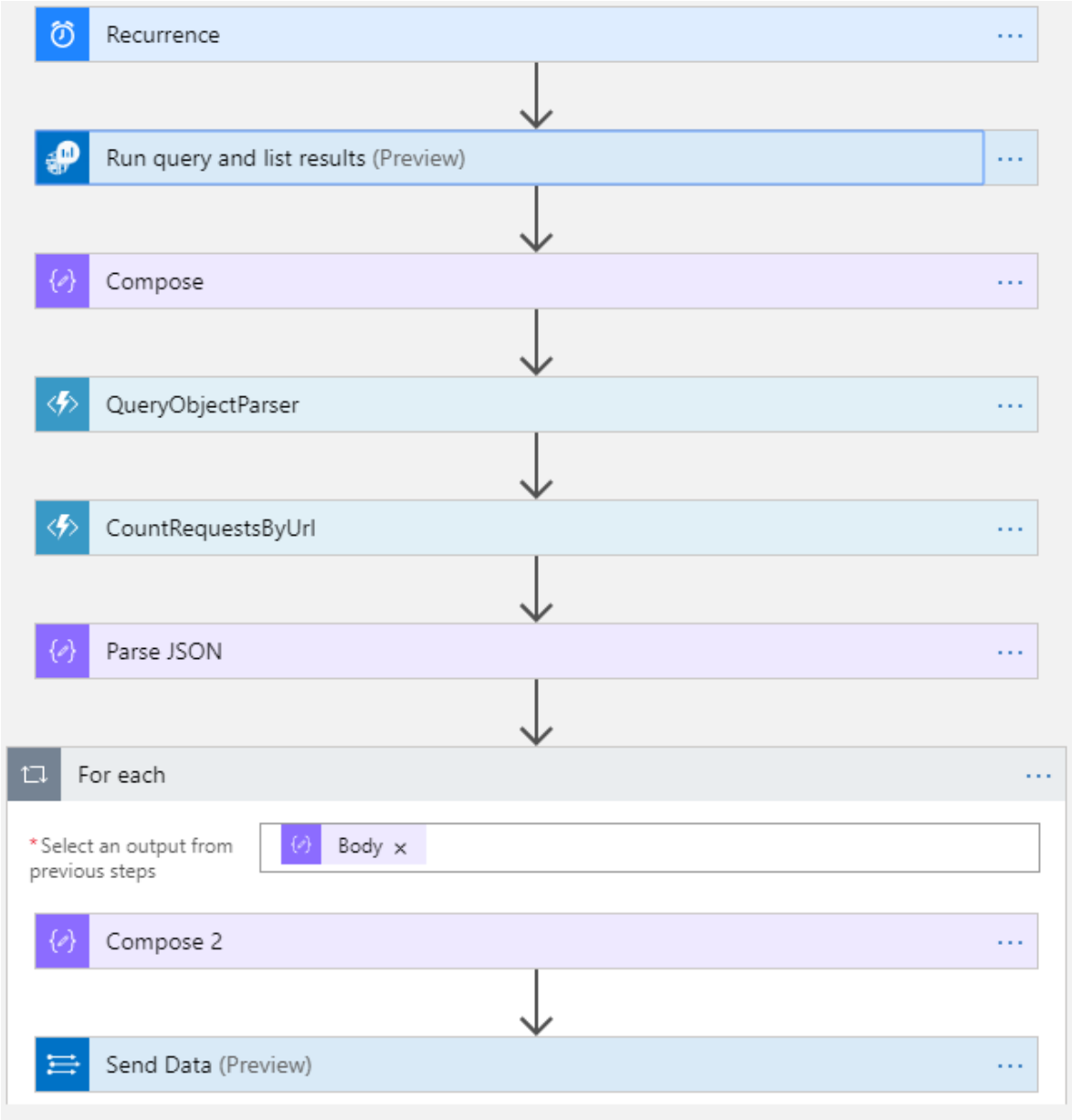


Figure 7. Final version of Logic App

5.2.3 Short reflection of current state and future possibilities

At the moment of writing this thesis, the project is finalized by publishing reports based on processed data. It brought a lot of new knowledge about source web site and its usage. The project is considered as very successful and its results were very valuable. Target organization started multiple new projects based on knowledge proceeded by this project and is very interested in taking more advantage of Microsoft Azure products.

6 Validating the Result of Proof-of-Concept

This chapter validates results of proof-of-concept from different perspectives taking care of all groups involved in implementation at some point. Software developers' perspective is described first as they were the most involved group at all phases. After that target customers' perspective is described as they are the end users of final products produced by proof-of-concept. The last but not the least perspective described is business owners' perspective as they are the ones, who's opinion about successfulness of proof-of-concept is major. Finally, all perspectives are summarized.

6.1 Perspective of Software Developers

The team formed for this Big Data Governance pilot project mostly consisted of software developers. CEO and CTO granted permission to involve all required human resources in this project. Actual implementation required only three software developers: me as main technical implementer and two supporting ones, who were responsible of validating technical correctness and processed data accuracy. Before this project they had a lot of experience in team working, so team communication during the project was not a challenge. Seamless team communication simplified planning and implementation phases very much. On the other hand, the team involved in the project had no experience in Big Data Governance approach and tools used for implementation were new for the team. Validating possibility of utilizing Big Data Governance approach and Big Data tools by small team without previous experience was one of main goals of this thesis.

Before starting pilot projects even meaning of term "Big Data" sounded strange and some of team members named it "just temporary hype" or "too complex to be implemented by small team". Of course, lack of experience in this area was a big challenge at the beginning. One of possible results of this pilot project was gaining knowledge, that Big Data Governance and Big Data overall is too complex or too immature technology to be utilized by small inexperienced team without big investments. A lot of tools that were used during implementation phases were new for the software developers and required learning curve could be very big obstacle.

However, the reality was quite opposite. Big Data Governance approach brought clearness to the projects highlighting areas of data governance which shouldn't be forgotten or underrated. Possibility to implement Data Governance theory in practice brought team to understanding, that most of DG areas were already well known and actively in use even without acknowledge that they are areas of DG. Understanding relations between these areas as a part of holistic data governance ideology was one of biggest achievements

gained by software development team involved in this pilot project. It gave the team deeper knowledge about Data Governance itself and Data Governance as a part of development process. This knowledge will definitely help software developers in another projects. Also, tools needed for processing Big Data were not that hard to learn as was expected before starting pilot projects. 5Vs of Big Data make it close to impossible to process real-time knowledge without tools specially designed for Big Data processing. However modern tools give possibility to utilize Big Data even by a team without experience in this area. Tools used for these projects have good documentation and very user-friendly and intuitive interface.

6.2 Perspective of Target Customer Groups

There were two separated pilot projects with different objectives and different outcomes. The first one was focusing on providing purely new value to organization's customers in form of new report available as a part of organization's service for its customers. In addition to that, actual subscriber of the project was organization's management. The second pilot project was focusing on bringing new knowledge to support development planning and organization's resource allocation. Therefore, there was two different target customer groups and their perspectives will be reviewed separately below.

6.2.1 Perspective of External Customer Group

External customer group consisted of all organization's customers who have access to organization's Owner Reporting Service. These customers are mainly real estate managers and real estate owners who use Owner Reporting Service to gain knowledge about their real estate properties. The new report, which was outcome of first pilot project was supposed to bring new knowledge in a simple and intuitive way, so most interesting information can be seen in a glance. Desired features of this new report were limited by knowledge collected from multiple sources and is shown in Appendix 2 (confidential). Not all these features were implemented during first iteration of this pilot project. List of implemented features is shown in Appendix 3 (confidential). Most of not implemented features were left on hold waiting for customers feedback for implemented features.

There were no satisfaction questionnaires for customers about this new report, so there is no statistical data about successfulness from customer perspective. Nevertheless, there were other ways to find it out. The outcome of second pilot project helped to collect data about report usage and compare it in time with other reports available to same customer group. This data showed that the report appeared in top 10 most used reports of Owner Report Service very soon after its publishing and stays there at the moment of writing this

thesis. In addition to that collected data also showed growth of popularity of related reports showing more detailed data about areas of this report. This knowledge ensured that expectation of usefulness of the report formed within pilot project was met. Also, there was a lot of customer feedback about this published report. Most of feedback contained gratitude about report informativeness and suggestions for further development. Many of suggestions were close to planned features which were not implemented during first iteration of the project. This was very inspiring as it gave organization's management confidence about project successfulness and correctness of future development plan.

6.2.2 Perspective of Internal Customer Group

Internal customer group consisted of target organization's management and owners. The outcome of first pilot project from their perspective was not product-customer relation. They were curious about knowledge that could be gained from outcome of second pilot project. Actual outcome was not a single report but rather processed data which can be queried to form different reports based on needs. Sample report was highlighting most used resources of web site and showed resources which were used rarely or not used at all. Without data processing automation this sample record could not be made. Organization's management was impressed by sample report and become more interested in gaining more knowledge from data which can be processed using similar approach. They suggested multiple new projects based on processed data. Also, organization's management suggested some new projects for processing other data using similar tools as were used in this pilot project. Overall, organization's management gave positive feedback about pilot project outcome and was enthusiastic about its further development.

6.3 Perspective of Business Owners

In the end of chapter 3.1. of this thesis are listed job titles of possible Data Governance participants. This list contains over forty different job titles and business owners were worried about possibilities to cover all Data Governance areas by a team that is many times smaller than that. Originally introduced in Table 4 list of most involved specialists consists only of three most involved participants data modeler, data architect and data steward plus several analysts covering all Data Governance areas. This approach showed its successfulness during pilot projects. It helped to divide areas of responsibility between pilot projects team members and request for additional help from specialists when needed. This gave business owners conclusion that Data Governance approach can be used within target organization regardless of lack of human resources compared to original list.

One of key factors interesting for business owners is ROI (return on investments) index. It is measured by comparing investments in project with income produced by results of the project. Actual ROI index for pilot projects was not calculated because neither of pilot project produce income which can be reliably separated from other target's organization income. However, implementation and support costs of both projects compared to gained results gave assumption that even ROI cannot be precisely calculated, actual value created was worth all investments. Implementation costs consisted only of human resources as there was no infrastructure investments. CEO together with CTO made everything possible so other organization's projects implementation is no affected by neither affect this pilot project. Support cost of first project is close to zero as new data processing is fully automated and it is using same resources that organization had before the project. Support costs of second pilot project were hard to calculate before implementation because the project uses Azure Cloud Services with pay-per-use payment model. Surprisingly, they stayed as low as 1-2 Euro per month, depending on amount of data to be processed. Compared to value produced by both projects the investments were extremely low and the projects can be described as very successful also from ROI perspective.

6.4 Summarizing all above

Both pilot projects were educational for software development team and gave a lot of new experience in Data Governance and possibilities to handle Big Data. Team formed during this project was satisfied with the results of their job and willing to adopt same techniques and work models in another projects. External and internal customers gave positive feedback on projects outcome and were interested in further development of these projects. They even suggested their vision of new features which was in line with planned development of the projects. Business owners were satisfied with work of software development team and project budget. CEO promised further support for both of projects. Summarizing all these perspectives results in awareness of total successfulness of proof-of-concept.

7 Conclusion and Further Development

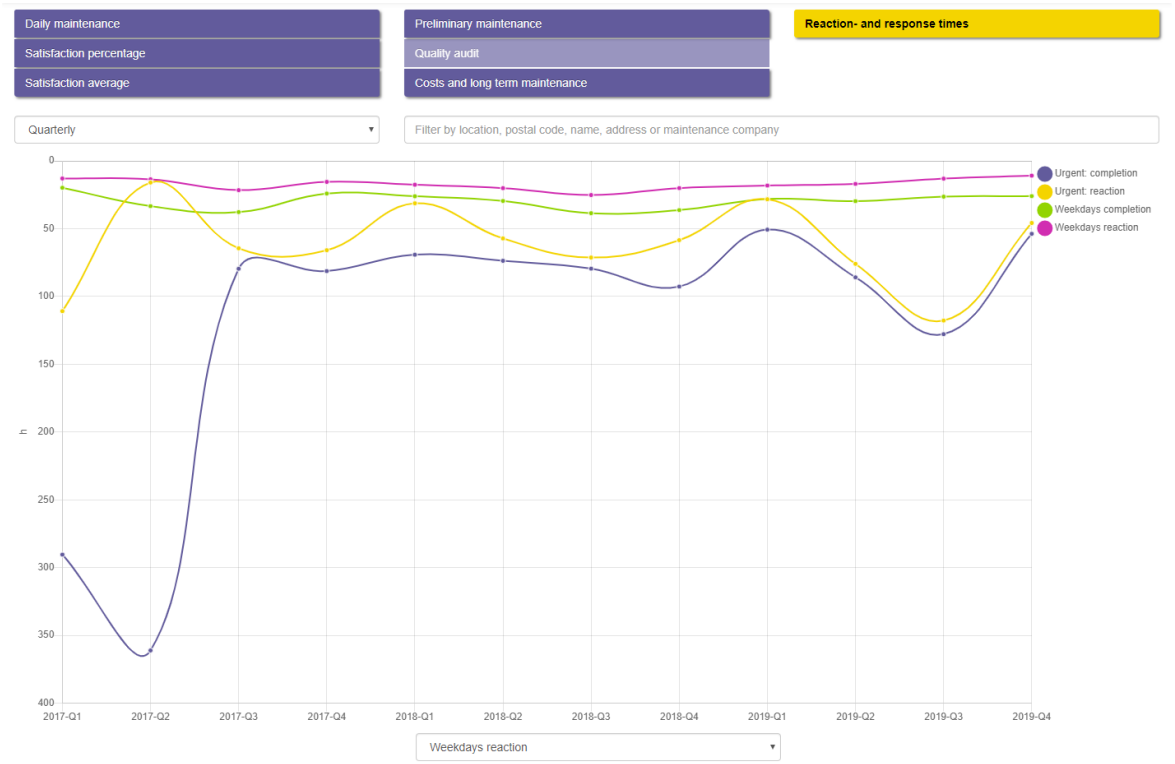
This chapter draws the bottom line for this research. First is shown customer value which was created during proof-of-concept phase. It ensures that this research had practical value for target organization. After that used methodology is reviewed in terms of suitability for described case and possible similar cases. It combines theoretical background with practical experience into framework which can be used for similar cases. Also, it analyses possibilities of Big Data and modern tools for manipulating with it. Finally, this chapter describes suggestions for further development within target organization and other possibilities to exploit knowledge gained during this research.

7.1 Created Customer Value

During the Proof-of-Concept phase there was two different pilot projects testing different tools for data processing and visualization. Both pilot projects were successfully finished, and they resulted in creating valuable outcome for target customer. As there were two pilot projects, their created value will be discussed separately.

7.1.1 Created Customer Value of First Pilot Project

First pilot project resulted in report available for target organization's customers as a part of Owner Reporting Service. This new report was named "Service level quality meters" and its main purpose is to show real estate managers and real estate owners all most interesting activities from their real estate objects. Figure 8 shows sample report from Service level quality meters. Actual data from existing buildings were used for this figure and names and addresses are overpainted with black color for privacy reasons.



Address /name of Object	2017-Q1	2017-Q2	2017-Q3	2017-Q4	2018-Q1	2018-Q2	2018-Q3	2018-Q4	2019-Q1	2019-Q2	2019-Q3	2019-Q4	Average
[redacted]			0	0.5			33.1	8.4	16.2	14.7	6.6	16.2	12
[redacted]	4	11.2	11.2	14.5	11.5	17.5	16.2	17.6	22.1	9	6	8.5	12.4
[redacted]	6.6	10.2	8.8	21.1	15.9	11.2	17.4	11.4	9.3	20.7	13.7	8.9	12.9
[redacted]	11.5	11.5	11.7	17.2	11.6	13.4	16.7	22.3	14.5	9.1	10.5	6.1	13
[redacted]	10.6	13.4	12.9	18.8	11.2	12.9	14.4	22.1	19.7	16.5	16.2	4.6	14.4
[redacted]				8.5	14.1	17.8	17.2	34.9	19.2	5.7	5.2	8.1	14.5
[redacted]	10	11.5	27.2	8.1	10	17.3	13	12.2	28.3	18.2	13.5	10	14.9
[redacted]	9	12.4	14.5	14.2	6.2	14.1	13.9	44.7	16.3	22.7	6.6	5.7	15
[redacted]				13.7	15.3	19.8	29.4	21.2	5.8	11.5	19.4	14.2	16.7
[redacted]	18.3	17.4	19.6	10.2	17.5	18.7	9.6	24	27.6	18.9	16.7	5.9	17
[redacted]	7	4.7	56.1	6.8	13.3	15.5	61.2	17	25.6	9.6	1.8	4.2	18.6
[redacted]	19.1	13.9	15.9	12.6	18.5	20.5	29.7	20	20.6	19.5	23.3	17.8	19.3
[redacted]	15.6	22.8	22.9	14.7	25.8	25.4	18.1	21	23.3	19	17.1	11.4	19.8
[redacted]	17.8	27.2	20.1	18.3	16.5	21.1	32.3	16.4	10.4	21.3	15.7	23.9	20.1
[redacted]	11.3	7.1	16.5	18.2	21.8	39.1	29.8	26.1	16	20.5	21.8	18	20.5
[redacted]	14.4	4.4	30.2	21.5	49.4	32.7	70.2	5.8	14.1	22.9	13	5.9	23.7
[redacted]	27.1	22.5	42.1	30.9	23.3	24.5	20.3	16.3	19.9	30	15.2	16	24
Average	**13**	**13.6**	**20.6**	**14.7**	**17.6**	**20.1**	**26**	**20.1**	**18.2**	**17**	**13.1**	**10.9**	**17**

Figure 8. Sample of created value of first pilot project

The report consists of seven areas grouped by subject of interest: daily maintenance; satisfaction percentage; satisfaction average; preliminary maintenance; quality and audit; costs and long-term maintenance; reaction and response times. Each of area consists of multiple meters. The meters can be switched off from chart to narrow focus only to interesting meters. The report is available for portfolio of real estate objects and for single object. If portfolio level report is selected, objects within the portfolio can be compared between each other by any meter within any area. Objects within portfolio can also be filtered by location, postal code, name, address or maintenance company. This simplifies focusing on interesting objects data without need to create new portfolio of objects. From

portfolio level it is easy to switch to another portfolio or navigate into single object for diving deeper into processed data.

Additionally, three report from processed data were brought to main pages of Owner Report Service portfolio and object level. Figure 9 shows sample of main page with three charts based on Service level quality meters on top of the page.

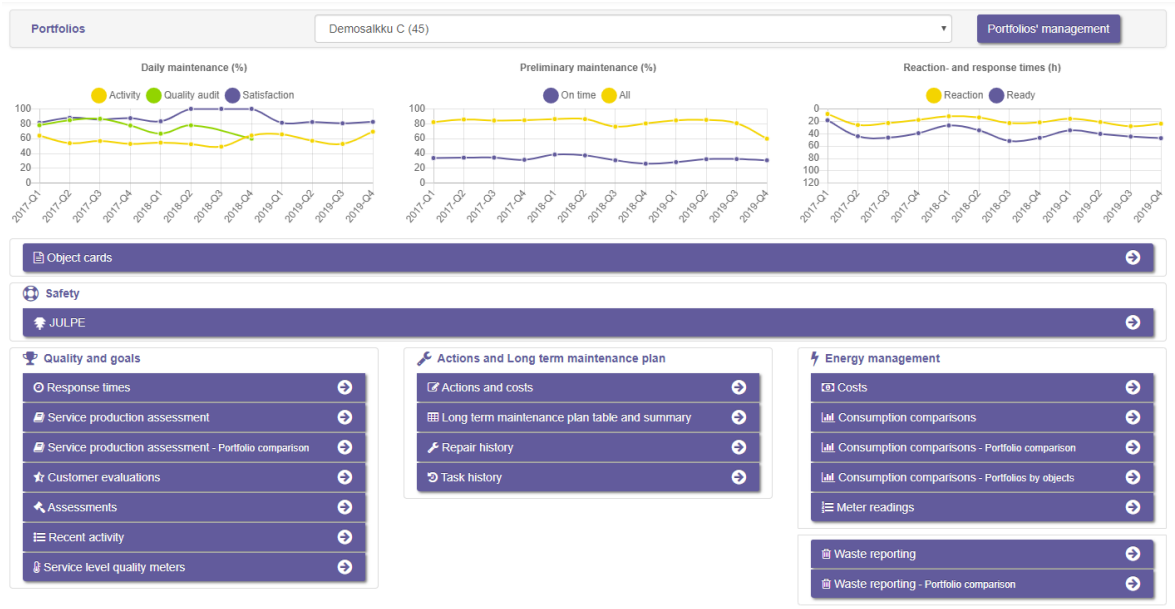


Figure 9. Main page of portfolio level Owner Report Service with charts

These charts are also clickable, and they navigate to related report in Service level quality meters. These three charts are showing data, which were reported by organizations customers as the most valuable as it allows to see possible issues in day-to-day activities.

Service quality report became one of top 10 most used reports of Owner Report Service and according to usage statistic its popularity keeps growing. In addition to that portfolio level main page, where three charts from the report are shown, nearly doubled page view counts within a week after Service quality report was published to customers. Some of customers also gave positive feedback for this newly created report and suggested new features for further development of the project. All these tells that actual value for target customers were created successfully.

7.1.2 Created Customer Value of Second Pilot Project

Customer of second pilot project was target organization's own management. The main reason for this project was to create knowledge about most used resources of target or-

ganization service and resources which were rarely used or not used at all. The project resulted in complex data processing job which calculates daily requests, combines them with all list of all available resources and stores calculated data in format which can be easily queried into table or chart reports. Figure 10 shows sample chart report which can be formed from processed data. It highlights most used resources and shows usage statistic by week for each of them. Actual resource names are overpainted with black color for target organization's confidential reasons.

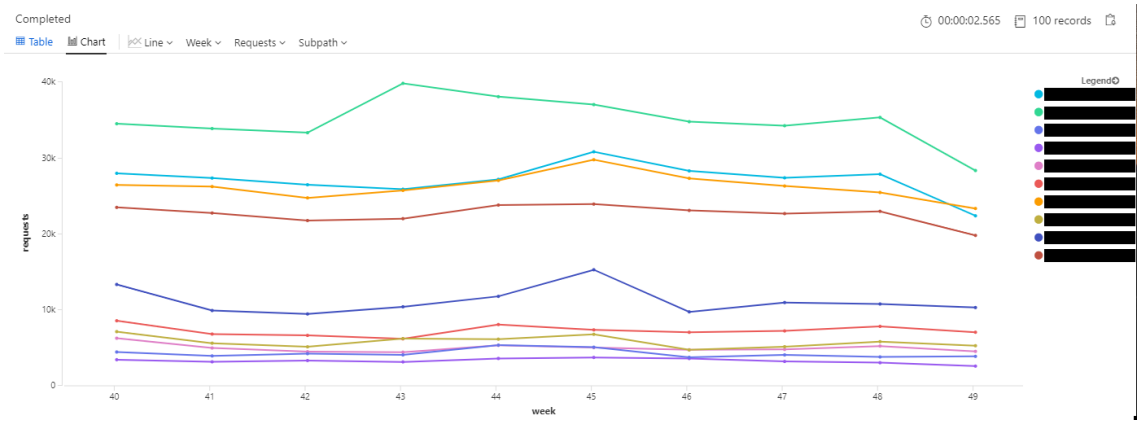


Figure 10. Most used resources grouped by week

Another sample report is shown in Figure 11. It lists resources which were not used since data processing was started. Here actual resource names are also overpainted with black color for target organization's confidential reasons. Figure 11 also shows how simple is querying of processed data.

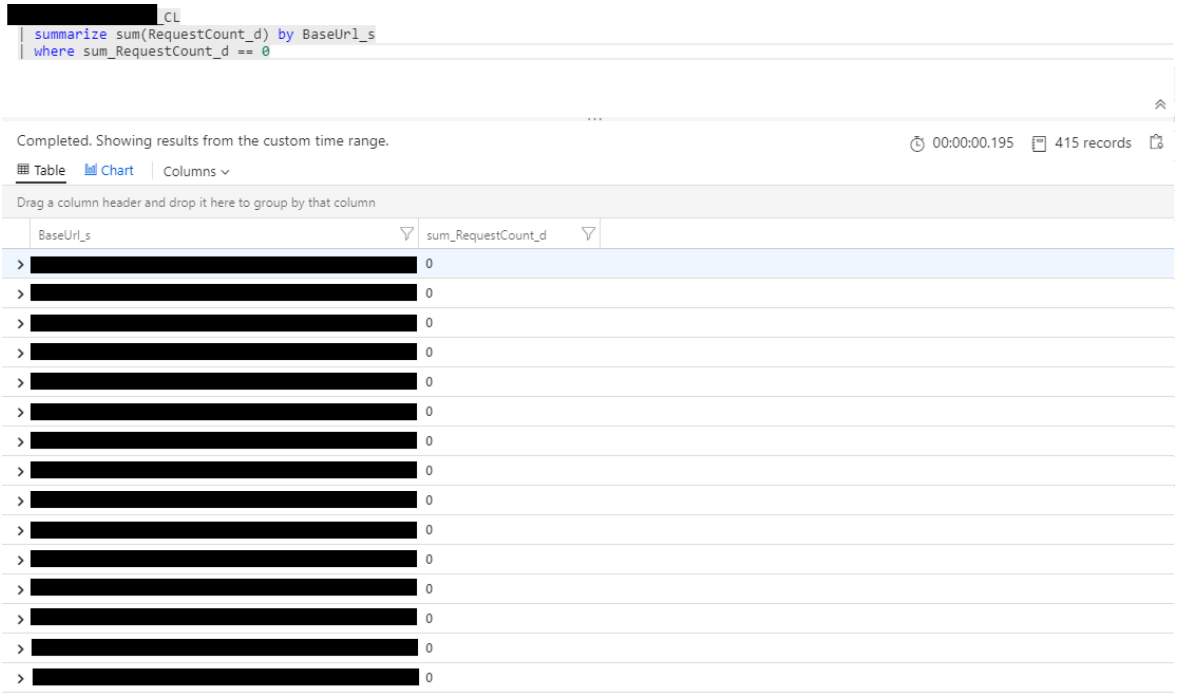


Figure 11. Resources without usage during data collection timespan

Both reports and any kind of other reports based on processed data can be pinned to Azure Portal Dashboard. Azure Portal Dashboard can be shared within target organization and using Auto refresh feature of dashboard it always shows up to date data. Reports can be also automatically sent via email to predefined audience, which makes it easy to catch up most interesting data in a very user-friendly way.

Target organization's management was impressed by demonstration of sample reports from processed data. CTO planned multiple new projects based on knowledge gained from processed data were CEO was interested in processing other data available for organization using same approach. Both were satisfied with outcome of pilot project and confirmed that it produces useful value for target organization.

7.2 Review of Used Methodology

During case study DAMA International's Data Management Framework was used as theoretical framework guiding planning and implementation of actual projects. Data which needed to be processed had signs of 5Vs of Big Data and suitable tools were brought into use. Importance of successful pilot project was acknowledged because Both Data Management Framework and Big Data were new for all participants. Below used methodology and its appearance in pilot projects is shown explicitly.

7.2.1 Data Management Framework

Data Governance was not new topic within target organization and there were existing data governance practices prior current research. During current research DAMA International's Data Management Framework was introduced to software development team and it was applied in proof-of-concept phase. It brought new viewpoint to data handling and software development in total.

Both pilot projects shared similar planning and implementation phases. They started from writing out desired result and collecting possible data sources. After that was turn to plan ETL (Extract Transform Load) process. These phases are tightly couped with each other because Transform phase is highly dependent on Load phase as data needs to be transformed to format most suitable for data loading. Extract phase in its turn is highly dependent on Transform phase as data must be extracted to the place and in format suitable for data transformation. Different areas of Data Management Framework took place during planning and implementation phases differently and below their effect in pilot projects are described in detail.

Data Architecture was in key role in both pilot projects in their all stages. Existing data architecture helped to discover available data sources, group them by subject and find data sources suitable for Extract phase of ETL. Planning ETL process itself required also planning of its data architecture because new ETL process creates new data flows. When projects were finished, their data architecture was added to Enterprise Data Architecture. This made newly generated data flows available for possible future projects.

Data Modelling and Design was also in very important role in both pilot projects in their all stages. Existing data models simplified limiting available data sources to only required ones and find relations between them. During Extract and Transform phases of ETL relations between data sources and knowledge how they can be merged was a key to success. Second pilot project required new Reference Data and produced new extracted data each requiring its own data model. Both pilot projects also produced new data in Transform phase so to accomplish this phase new data models were designed.

Metadata management took place in exactly same phases as Data Modelling and Design adding descriptive information about data. Metadata kept up to date helped to discover data sources and estimate their suitability and accuracy The only difference was that when new data models for new data were created prior to actual existence of new data, Metadata was written out after newly created data was validated.

Reference and Master Data was most important in Extract phase and preparation for Transform phase of ETL. It added clearance to these phases separating continuously changing Master Data from nearly static Reference Data. In first pilot project Master Data consisted of all data sources required for desired result where the list of collected metrics played role of Reference Data. In second pilot project Master Data was real-time logs from web application and Reference Data was formed from all available resources within the web application.

Data Storage and Operations was in key role in all phases where new data was generated. For first pilot project it was Load phase of ETL where transformed data was loaded back to target organization's database. For second pilot project it was Extract and Load phases. In Extract phase reference data was firstly saved locally and then sent to the same cloud service where Master Data was collected. In Load phase of second pilot project transformed data was also stored in same cloud service.

Data Integration and Interoperability was tightly couped with Data Architecture, Data Modelling and Design in all phases checking that data is available for integration. Also, it affected Data Storage and Operations by bringing clearance to possible locations of processed data. Data Integration and Interoperability helped to ensure that data to be processed can be integrated between each other and that processed data is available for further operating.

Data Warehousing and Business Intelligence took place in exactly same phases as Data Integration and Interoperability although it was not so visible. None of pilot projects required new data warehouses or data marts. Result of first pilot project enriched business intelligence experience of target organization's customers. Result of second pilot project enriched business intelligence experience of target organization's management with possibility to broaden BI experience by creating new analyses on processed data.

Document and Content Management was no introduced during pilot projects as none of them was manipulating nor producing any kind of documents.

Data Security played very important role in all phases in both pilot projects. As both projects retrieved existing data and produced new data, importance of data security and data privacy was at very high priority level. In first pilot project Data Security was carried out by implementing ETL-process using target organization's in-house environment, which is continuously validated to be secure by internal audits. Data privacy of first pilot project

was handled by putting projects results in same permission checking framework as other parts of Owner Report Service. Data Security and Data Privacy of second project was carried out by placing whole ETL-process and project results into cloud service access to which is limited to authorized target organization's personnel only. All activities and all handled data were continuously audited by target organization's DPO (Data Protection Officer) to comply with governance regulations and target organization's security and privacy policy.

Data Quality was mostly visible during Extract and Transform phases of both pilot projects. First data to be extracted was validated to be high-quality data because it would be close to impossible to gain new knowledge from poor quality data. During Transform phases all processes which could be automatically tested to produce correct data were automatically tested on sample data. These processes which could not be automatically tested were audited by multiple specialists to ensure correctness of these processes.

7.2.2 Big Data

Big Data adds complexity to data processing and processing Big Data into information could be a challenge for small-sized organization or a team without previous experience in that area. Traditionally, Big Data is considered as something that is hard to manipulate with or something that requires expensive environment to deal with it. Learning gap and big investments could be reasons for lagging in technology utilizing. During case study both assumptions were refuted.

Cloud based tools for Big Data processing are evolving rapidly. They have user-friendly user interface and allow to build data processing solutions in just few clicks at its best. Ready to use blocks provide versatile options for retrieving, processing and delivering processed data. They hide most of technical setting making it easy to focus only on data processing leaving connections between blocks to be handled automatically in background. In cases were some data processing activities require complex logical operations and could not be accomplished using built-in tools, modern cloud platforms allow to write your own data processing blocks using any of most popular modern programming language. These custom blocks can be connected with other blocks within cloud solution or used as standalone tiny apps splitting data processing or any other repetitive task into pieces small enough for handling and testing. Cloud based tools also provide wide assortment of connections for retrieving and delivering data. These connections support all data storage options within cloud ecosystem and various options for interact with external data storage options. If existing connection options are not enough, modern cloud-based tools allow to create your own custom connections, expanding possibilities of these tools even more.

When it comes to the investment comparison, modern cloud-based tools are very competitive choice. Solutions built using them are handled as self hosted apps which means that they don't need physical nor even virtual environment setup. Creator of Big Data solution don't need to worry about operating systems, network rules, memory allocation etc. aspects which are important when building traditional physical or virtual solution environment. This lowers investments in environment and reduces amount of required support functions making possible to create and maintain Big Data solutions using very limited monetary and human resources. In addition to that, cloud-based solutions are charged for actual usage only. That means that there are no costs for any standby time, which is common when speaking about traditional physical or virtual servers. If data processing solution is not running continuously but is triggered by an action or schedule, pay-per-usage model can be notably cheaper than other alternatives.

Every time speaking about Big Data, cloud environments or their combination data security and data privacy issues arise. Governance regulations and enterprise data security policies could add limitations on processing data in the cloud. These regulations and policies are obligatory and must be carried out during all phases of Big Data solution development. Fortunately, modern cloud-based environment providers offer possibilities to build solutions in a manner where all data security and data privacy obligations are met.

As a conclusion it could be admitted that modern cloud-based tools make Big Data processing possible without deep knowledge in this area, in cost efficient way and with respect to data security.

7.2.3 Big Data Governance in small-sized organization

Each sub area of Data Management Framework consists of multiple activities each having different list of possible participants. Because of small size of target organization Data Management Framework areas were grouped by most involved participants. They were reduced to data architect, data modeller, data steward and several analysts. Case study showed, that in situations, where available human resources are strictly limited by size of organization, this short list of participants covers all areas of Data Management Framework. Furthermore, case study showed, that roles of data architect and data modeller are so tightly couped that in special conditions same person could possibly act as both roles. Data steward role showed to be so important that it could not be removed nor replaced by any other involved specialist. Analysts were involved only in situations where their knowledge, finesse or authority were necessary for projects development.

Due to its nature Big Data makes data governance even more big challenge. However, case study showed that Data Management Framework is applicable also in environments where data to be processed has features of Big Data. During planning and implementation phases of both projects of current case study highlighted multiple similarities although technical solutions for the projects were different. This observation evolved insight that these similarities combined with theoretical Data Management Framework could form practical model for developing data processing solutions caring for all data management activities. Figure 12 shows framework which is applicable to the both projects of case study.

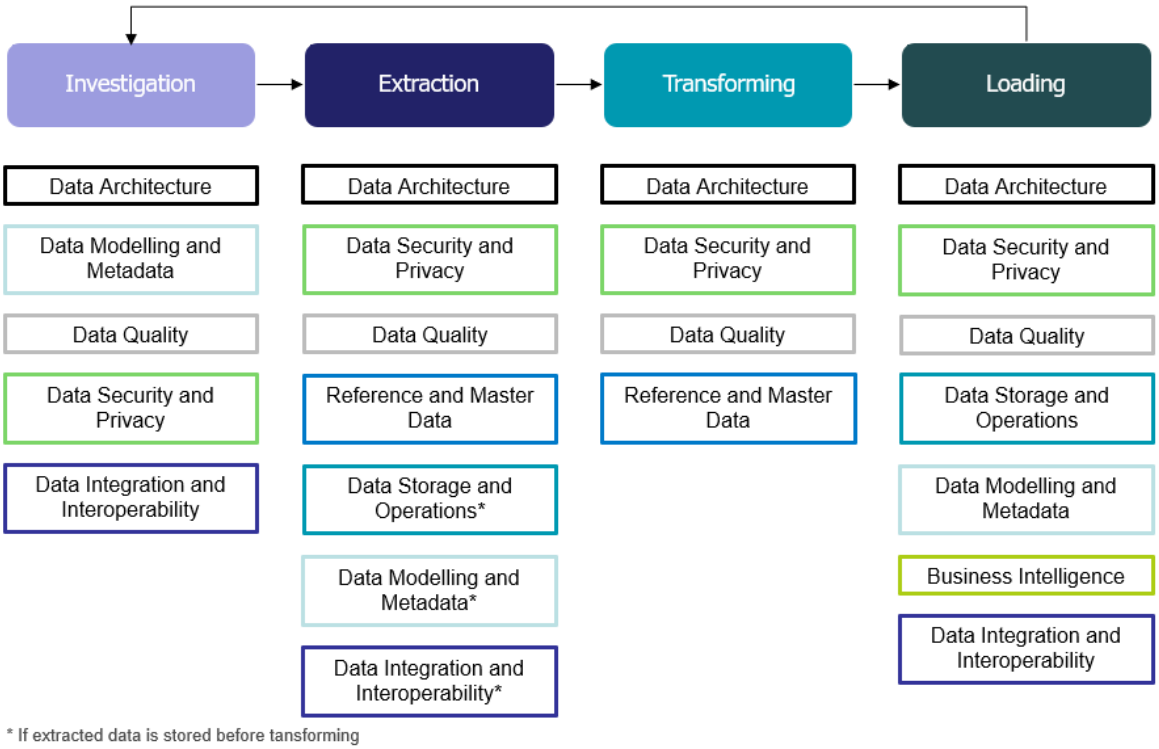


Figure 12. Data Governance Framework for Data Processing

Data Governance Framework for Data Processing extends ETL (Extract Transform Load) process with new step Investigation. This step is initial for any data processing project because no data extraction can be started without previous investigation on existing data and possible sources for new data. ETL process with preparation for it seems to be key process for a project which purpose is to create insights from existing data. Arrow back from Loading phase to Investigation phase describes that this process is continuous and iterative. Each time new insights are gained from data new round of this process can be started to process these insights furthermore.

As can be seen in Figure 12, original areas of Data Management Framework were also adjusted based on pilot projects of case study. Data Modelling and Design area was merged with Metadata area into common Data Modelling and Metadata. Although Data Modelling and Metadata are different areas with their own targets, during the case study was shown, that they both can and should be accomplished at the stage when new data is created to describe that data. Data Warehousing and Business Intelligence was split into separated areas as it was shown that Business Intelligence can be accomplished without Data Warehousing. Data Warehousing was not introduced during case study as there was no need for new data warehouses. Possibly activities under Data Warehousing could be covered by combination of Data Architecture, Data Storage and Operations, Data Integration and Interoperability, Data Security and Privacy. Document and Content Management was not introduced during case study and therefore it is not present in Figure 12. Possibly its activities take place in Investigation, Extraction and Loading phases of Data Governance Framework for Data Processing. Data Security was extended by Data Privacy. Second one misses from titles of original Data Management Framework, but it is described as one of activities within Data Security area. Data Privacy is very important part of Data Governance. Modern governance regulations about human data privacy all around the World just establish this importance. Therefore, Data Privacy was brought to the titles of Data Governance Framework for Data Processing.

7.3 Suggestions for Further Development

Target organization has clear vision for further development of the projects which were launched during case study. Based on new insights and feedback from customers both projects will keep evolving and generating new insights from processed data. Another area where target organization could exploit this case study is bringing experience of Data Governance and Big Data into other projects. This could be done by sharing experience of team involved in case study in organizational meetings and by involving new organization's personnel into projects with elements of Data Governance and Big Data. Target organization's management is very interested in continuous improvement of Data Governance policies and new possibilities of Big Data.

Data Governance Framework for Data Processing provided in Figure 12 is formed based only on two projects it could not be treated as final framework and ready to use guideline. It needs to be tested on multiple projects and possibly adjusted to gain its final form. This framework could be possibly used for new iterations of case study projects and for other projects within target organization. Also, this framework needs to be tested on other small sized organizations by other teams to be validated as working guide for Data Management. Because neither of projects within case study contained Document and Content

Management, framework need to be tested on projects which includes this area of Data Management to enrich the framework by it. Data, its forms, regulations about it, tools for data processing and everything other related to it is changing continuously, so Data Governance Framework for Data Processing needs to be continuously improved to reflect up-to-date state of Data Governance.

Discussion

Research Questions Results and New Research Questions

The main objective of this thesis was to test possibility of creating value from Big Data using Data Governance approach by small team within small-sized organization. Original research questions were:

- Can small-sized organization create value from Big Data without previous experience in that area?
- Can Big Data Governance be handled with respect in situations, where human resources are strictly limited?

The short answer is “yes and yes”. Longer answer question by question is described below. After it other findings are discussed and new research questions are presented.

Big Data is a huge topic which is still surrounded with hype and assumptions of complexity. During case study it was proved that modern cloud-based tools for Big Data processing have very user-friendly interface and allow to create Big Data processing solutions even without deep knowledge in technical aspects of Big Data processing. This lower learning gap and technical skills requirements. Solutions made by these cloud-based tools don’t need physical or virtual computer environment installations and are charged per usage. This lower required installation investments and support costs and makes processing of Big Data available for any size of organization. Actual support costs of Big Data processing in case study projects was as low as a cup of coffee per month.

Big Data Management is a multidimensional discipline each sub-area of which consists of multiple activities and therefore involves a lot of possible participants. During case study was shown that this topic can be handled by cooperation of Data Architect and Data Steward who involve other specialists in cases where their own knowledge is not enough. To make it possible support from organization’s management is required. Because Data Management is more about organizational global practices than about technical side of its activities, it is very important that everybody within target organization is aware of Data Management policies. Data Management should not be a single time task nor even repeated task but rather a guidance for any processes which contains any kind of data.

Theoretical framework described in chapter 3 did not provide ready to use guide for Big Data Governance practices for small-sized organizations. However, after case study based on accomplished projects Data Governance Framework for Data Processing was formed and provided in Figure 12 in chapter 7. Despite technical difference and different target customer of the projects, there were similarities in development phases and Data

Management activities within these phases. Although this new framework is not tested more broadly it could be used as a data management guide for other data processing projects.

One of topic for new research could be comparison of providers of Big Data processing tools and environments. During current case study only one set of tools from one of most popular Big Data cloud-solution providers was tested. It makes impossible to form any assumptions about its pros and cons compared to other cloud-solution providers or to in-house hosted Big Data environment. This comparison based on practical solution could be very useful in many cases.

Freshly formed Data Governance Framework for Data Processing needs to be tested and validated by other projects and within other organizations. Possibilities to use same framework in medium-sized and big organizations can also be evaluated. Wider research about correctness of this framework could adjust it to the form of ready to use guide for data processing projects. This guide can be very useful for focusing on most important areas of Data Management at any stage of data processing project.

Personal evaluating

This case study was very valuable experience which made me to learn a lot of new things. First of all, Data Governance itself was quite new topic to me. Before this thesis I was involved in multiple activities related to Data Governance, but I was not aware about their deeper relations and didn't have clear picture about all activities which are part of Data Governance. This thesis cleared Data Governance activities, their relations between each other and showed practical implementation of them. Also processing of Big Data was purely new experience for me. Before writing this thesis, I read a lot about Big Data and had some basic knowledge about its theoretical side and close to no knowledge about Big Data processing options and tools. During this thesis I became aware about wide range of tools for Big Data processing available. Furthermore, I tested one of these tools and confirmed that it is simple to use even without previous experience in Big Data. Due to nature of this thesis I managed to accomplish two pilot projects during it. This made possible to find similarities between them and form base for new framework which possibly could be used widely for taking care of all areas of Data Management.

References

Alley, G. 2019. What is Big Data Architecture? URL: https://dzone.com/articles/what-is-big-data-architecture Accessed: 11 September 2019

An Introduction to Integration and Interoperability 2003. URL: https://www.im-magic.com/eLibrary/ARCHIVES/GENERAL/SUNGARD/S030200I.pdf Accessed: 10 September 2019

Aunimo, L., Alamaki, A. V., Ketamo, H. 2019 Big Data Governance in Agile and Data-Driven Soft-ware Development: A Market Entry Case in the Educational Industry. Big Data Governance and Perspectives in Knowledge Management, pp. 179-199.

BI and Data Warehousing: Do You Need a Data Warehouse Anymore? 2019. URL: https://panoply.io/data-warehouse-guide/bi-and-data-warehousing/ Accessed: 13 September 2019

Big data analytics 2019. URL: https://cloud.google.com/solutions/big-data/ Accessed: 15 September 2019

Big Data Architectures 2018. URL: https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/ Accessed: 11 September 2019

Big data products 2019. URL: https://cloud.google.com/products/big-data/ Accessed: 15 September 2019

Cai, L. and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14, p.2.

Castanedo, F. & Gidley, S. 2017. Understanding metadata: Create the foundation for a scalable data architecture. O'Reilly Media, Inc.

DAMA International 2017. DAMA-DMBOK: data management body of knowledge. 2nd ed. Technics Publications.

Data Lakes and Analytics on AWS 2019. URL: https://aws.amazon.com/big-data/data-lakes-and-analytics/?nc1=h_ls Accessed: 15 September 2019

Denzin, N. K. & Lincoln, Y. S. 2000. Handbook of qualitative research. 2nd ed. Thousand Oaks. Sage Publications Inc.

El Bassiti, L. 2019. Big Data, Semantics, and Policy Making: How Can Data Dynamics Lead to Wiser Governance? Big Data Governance and Perspectives in Knowledge Management, pp. 154-178.

English, L. 2019. The Role of the Pilot Project in Effective Organizational Change. Lean Enterprise Institute

Gaille, B., 2018. 15 Microsoft Azure Advantages and Disadvantages. URL: https://brandongaille.com/15-microsoft-azure-advantages-and-disadvantages/ Accessed: 15 September 2019

Gandham, M. 2019. What are the pros and cons of Google Cloud Platform? URL: https://www.quora.com/What-are-the-pros-and-cons-of-Google-Cloud-Platform Accessed: 15 September 2019

Gartner 2019. Magic Quadrant for Data Science and Machine Learning Platforms. URL: https://www.gartner.com/doc/reprints?id=1-65WC0O1&ct=190128&st=sb Accessed: 15 September 2019

Gartner 2019. Magic Quadrant for Cloud infrastructure as a Service. URL: https://www.gartner.com/en/documents/3947472/magic-quadrant-for-cloud-infrastructure-as-a-service-wor Accessed: 15 September 2019

Gupta, Y. 2018. Importance of doing a Pilot Project before Full Scale Automation Tool Roll Out. URL: https://www.softwaretestinggenius.com/importance-of-doing-a-pilot-project-before-full-scale-automation-tool-roll-out/ Accessed: 16 September 2019

Huda, M. M., Hayun, D. R. L., Zhin, M. 2015. Data Modeling for Big Data. Informatics Engineering, ITS, Surabaya, Indonesia

Ishwarappa & Anuradha J. 2015. A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology. Procedia Computer Science, Vol. 48, pp. 319-324.

Ladley, J. 2012. Data Governance: How to Design, Deploy and Sustain an Effective Data Governance Program. Morgan Kaufmann.

Lonoff Schiff, J. 2013. 14 Things You Need to Know About Data Storage Management. URL: https://www.cio.com/article/2382585/14-things-you-need-to-know-about-data-storage-management.html Accessed: 12 September 2019

Marr, B. 2019. How Can Small Businesses Use Big Data? Here Are 6 Practical Examples. URL: https://www.bernardmarr.com/default.asp?contentID=1442 Accessed: 13 December 2019

McGilvray, D., Thomas, G. 2008. Definitions of Data Categories. Morgan Kaufmann Publishers. Elsevier Inc.

McLeod, S. A. 2017. Qualitative vs. quantitative. URL: https://www.simplypsychology.org/qualitative-quantitative.html Accessed: 11 March 2018

Nero, T. 2018. Metadata Management in Big Data. URL: https://www.cue-logic.com/blog/metadata-management-in-big-data-systems Accessed: 10 September 2019

Pomerantz, J. 2015. Metadata. The MIT Press.

Pros and Cons of using Amazon Cloud Services 2018. URL: https://www.attunix.com/insights/pros-cons-using-amazon-cloud-services/ Accessed: 10 September 2019

Roberts, B. 2017. Integration vs Interoperability: What's the Difference? URL: https://blog.sisfirst.com/integration-v-interoperability-what-is-the-difference Accessed: 13 September 2019

Roland van Teijlingen, E. & Hundley, V. 2002. The Importance of Pilot Studies. Nursing Standard: Official Newspaper of the Royal College of Nursing, Vol. 16, Issue 40, pp. 33-36.

Sarsfield, S. 2009. The Data Governance Imperative. IT Governance Publishing. Cambridgeshire.

Saunders, M., Lewis, P., Thornhill, A. 2016. Research methods for business students. 7th ed. Pearson Education Ltd. Harlow.

Schmarzo, B. 2018. Importance of Metadata in a Big Data World. URL: https://www.data-sciencecentral.com/profiles/blogs/importance-of-metadata-in-a-big-data-world Accessed: 10 September 2019

Simsion, G., Milton, S. K., Shanks, G. 2012. Data modeling: Description or design? Information & Management. Vol. 49, Issues 3-4, pp. 151-163.

Spacey, J. 2016. Reference Data vs Master Data. URL: https://simplicable.com/new/reference-data-vs-master-data Accessed: 13 September 2019

Talhaa, M., El Kalam, A. A., Elmarzouqi, N. 2019. Big Data: Trade-off between Data Quality and Data Security. Procedia Computer Science, Vol. 151, pp. 916-922.

Tankard, C. 2012. Big data security. Network Security, Vol. 2012, Issue 7, pp. 5-8.

Tawalbeh, A. & Saldamli, G. 2019. Reconsidering Big Data Security and Privacy in Cloud and Mobile Cloud Systems. Journal of King Saud University - Computer and Information Sciences.

Templafy 2019. ocument management vs. content management: What is the difference? URL: https://www.templafy.com/blog/document-management-vs-content-management-what-is-the-difference/ Accessed: 13 September 2019

Top 7 Advantages of Using Azure Cloud Services 2018. URL: https://www.attunix.com/insights/top-7-advantages-using-azure-cloud-services/ Accessed 15 September 2019

Turner, R. 2005. The Role of Pilot Studies in Reducing Risk on Projects and Programmes. International Journal of Project Management, Vol. 23, Issue 1, pp. 1-6.

Verhoef, P. C., Kooge, E., Walk, N. 2016. Creating Value with Big Data Analytics: Making Smarter Marketing Decisions. 1st ed. Routledge. London and New York.

Wang, U. 2016. How Google is using big data to protect the environment. URL: https://www.theguardian.com/sustainable-business/2016/oct/12/google-environmental-sustainability-data-kate-brandt Accessed: 14 September 2019

Washington, E. 2019. Why Data Governance is Crucial for Big Data Environments. URL: https://www.readitquik.com/articles/data/why-data-governance-is-crucial-for-big-data-environments/ Accessed: 14 December 2019

What Is the Difference between a Trial and a Pilot? 2019. URL: https://www.apm.org.uk/resources/find-a-resource/what-is-the-difference-between-a-trial-and-a-pilot/ Accessed: 16 September 2019

Zachman, J. A. 2008. Zachman Framework. Zachman International, Inc.

Zbrodoff, S. 2012. Pilot projects--making innovations and new concepts fly. URL: https://www.pmi.org/learning/library/pilot-projects-innovations-new-concepts-6043 Accessed: 16 September 2019

Zhang, L. 2018. A Comparison of Data Modelling Methods for Big Data. URL: https://dzone.com/articles/a-comparison-of-data-modeling-methods-for-big-data Accessed: 11 September 2019

Zhou, Z., Huang, D. 2012. Efficient and Secure Data Storage Operations for Mobile Cloud Computing. Arizona State University

Appendices

Appendix 1. Screenshots of Implementation Phase (confidential)

Appendix 2. List of desired features of first pilot project (confidential)

Appendix 3. List of implemented features of first pilot project (confidential)

Appendix 4. Published news about project

URL: https://palvelu.fimx.fi/news/laatumittarien_trendigraafit/fi Accessed 01.09.2019

URL: https://palvelu.fimx.fi/news/mistalaatumittaritkoostuu/fi Accessed 01.09.2019