Aalto University
School of Business

Review of Modern Business Intelligence and Analytics in 2015: How to Tame the Big Data in Practice?

Case study - What kind of modern business intelligence and analytics strategy to choose?

MSc program in Information and Service Management

Master's thesis

Samu Kulin

2015

Department of Information and Service Economy
Aalto University
School of Business

Aalto University
School of Business

Aalto University, P.O. BOX 11000, 00076
AALTO
www.aalto.fi

Abstract of master's thesis

Author Samu Kulin

Title of thesis Review of modern business intelligence and analytics in 2015: How to tame the big data in practice / Case study – What kind of modern business intelligence and analytics strategy to choose?

Degree Master of Science in Economics and Business Administration

Degree programme Information and Service Economy

Thesis advisor(s) Matti Rossi

Year of approval 2015 **Number of pages** 84 **Language** English

Abstract

The objective of this study was to find out the state of art architecture of modern business intelligence and analytics. Furthermore the status quo of business intelligence and analytics' architecture in an anonymous case company was examined. Based on these findings a future strategy was designed to guide the case company towards a better business intelligence and analytics environment. This objective was selected due to an increasing interest on big data topic. Thus the understanding on how to move on from traditional business intelligence practices to modern ones and what are the available options were seen as the key questions to be solved in order to gain competitive advantage for any company in near future.

The study was conducted as a qualitative single-case study. The case study included two parts: an analytics maturity assessment, and an analysis of business intelligence and analytics' architecture. The survey included over 30 questions and was sent to 25 analysts and other individuals who were using a significant time to deal with or read financial reports like for example managers. The architecture analysis was conducted by gathering relevant information on high level. Furthermore a big picture was drawn to illustrate the architecture. The two parts combined were used to construct the actual current maturity level of business intelligence and analytics in the case company. Three theoretical frameworks were used: first framework regarding the architecture, second framework regarding the maturity level and third framework regarding reporting tools. The first higher level framework consisted of the modern data warehouse architecture and Hadoop solution from D'Antoni and Lopez (2014). The second framework included the analytics maturity assessment from the data warehouse institute (2015). Finally the third framework analyzed the advanced analytics tools from Sallam et al. (2015).

The findings of this study suggest that modern business intelligence and analytics solution can include both data warehouse and Hadoop components. These two components are not mutually exclusive. Instead Hadoop is actually augmenting data warehouse to another level. This thesis shows how companies can evaluate their current maturity level and design a future strategy by benchmarking their own actions against the state of art solution. To keep up with the fast pace of development, research must be continuous. Therefore in future for example a study regarding a detailed path of implementing Hadoop would be a great addition to this field.

Keywords modern business intelligence, data warehouse, big data, Hadoop, strategy

ABBREVIATIONS

BI&A	Business Intelligence and Analytics
DB	Database
DBMS	Database Management System
DW	Data Warehouse
DM	Data Mart or Data Model
EDM	Enterprise Data Model
EDW	Enterprise Data Warehouse
ERP	Enterprise Resource Planning
ETL	Extract, Transform and Load
GPS	Global Positioning System
HDFS	Hadoop Distributed File System
ICT	Information and Communication Technology
IT	Information Technology
KPI	Key Performance Indicators
MIT	Massachusetts Institute of Technology
MPP	Massively Parallel Processors
ODS	Operational Data Store
OLAP	Online Analytical Processing
OLTP	Online Transaction Processing
PDF	Portable Document Format
SAP	System Analysis and Program Development (software)
SAS	Statistical Analysis System (software)
SQL	Standard Query Language
TDWI	The Data Warehouse Institute

Acknowledgements

I owe a debt of gratitude to my dear wife Anni for her patience during the moments when I was absent due to thesis work. I love you. I would also like to thank my friends and relatives for their support throughout the spring.

Writing this thesis has been both a painstaking and an enjoyable process. Special thanks belong to my thesis advisors Professor Rossi M. and Assistant Frantsev A. for their time to give me guiding comments and practical assistance along the thesis writing process and seminar arrangements. Also opponents Oittinen J. and Autero T. deserve a mention for their valuable feedback and fine-tuning comments at the last mile of my writing process.

Final thanks goes to the Dropbox service for keeping excellent care of my files and making them available anywhere and anytime.

Thank you to all of you.

June 2nd, 2015

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION

Our past tells us that the development regarding industrial revolutions like steam power and electricity lasts roughly one hundred years. Any new technology for general purpose can accelerate labor productivity in three different stages. First, for example efficiency in the production of computers and mobile phones is increased by new technology. Second, labor productivity is increasing when computers and mobile phones are used in other industries as well. The third phase is taking its place when companies are changing their ways of doing things e.g. outsourcing their work tasks via Internet. (Pohjola 2010, 158) Only by increasing labor productivity we can reach the next level in standard of living due to fact that natural resources and population of mankind is limited.

In 2015, third industrial revolution, the era of information and communication technology has lasted forty years and the earlier industrial revolutions' benefits of steam power and electricity are almost fully utilized. The seeds of ICT-era were sown back in the mid-1970s and since then our world, economy and business world have been evolving and our mankind has taken some major steps towards better standard of living with the help of new information and communication technology. Since 1995 one of the main drivers has been the Internet (Pohjola 2010, 155.) It has been said that history repeats itself and if we rely on that phrase we could place ourselves somewhere in the middle of ICT development or third industrial revolution. Matti Pohjola (2010, 158) claims that at the moment we are about to enter the third step of our third industrial revolution. This means companies are changing their ways of doing things, with data. McKinsey Global Institute (MGI 2011, 2) is on the same page with Pohjola; MGI suggests that we are on the cusp of a tremendous wave of innovation.

One major difference compared to earlier breakthroughs is our brain capacity. We are not anymore, significantly, increasing our working power except for computing. Instead we are able to take our intelligence to another level. The amount of data we are handling nowadays is just magnificent and almost out of control, approximately 2.5 Exabyte (1 EB = 10^{18} = 10 million library floors of academic journals (UC Berkley, 2003)) of data were created each day in 2012 and this number is doubling almost every 40 months (McAfee & Brynjolfsson, 2012). With efficient use of this big aata we are able to make better business decisions including market analysis and more accurate predictions (Johnson 2012, 53).

Simply put, big data is diverse information that companies use for their own business purposes.

The more companies characterize themselves as data-driven, the better they perform on objective measures of financial and operational results (McAfee & Brynjolfsson, 2012). According to MIT Center for digital business, companies that make data-driven decisions are 5% more productive and yield on average 6% more in profits. Based on these claims and studies I agree with Barton & Court (2012, 81) that one way to make better data-driven decisions is to manage big data and treat it as a valuable and strategic asset. However we should not forget the criticism. According to McAfee and Brynjolfsson (2012, 63) not everyone is embracing data-driven decision making. Can we rely on these studies or shall we treat the concept of big data as if it was an overvalued hype? One way to get involved with big data is to develop current business intelligence and analytics architecture. Many trends in business intelligence and analytics are relevant for companies looking to become more mature in their analytics efforts. Halper and Stodder (2014, 6) identify eleven trends. These include:

1. Ease of use
2. The democratization and consumerism of analytics
3. Business analysts using more advanced techniques
4. Newer kinds of analytics
5. Operationalizing analytics
6. Big data
7. New development methods
8. Open source
9. The cloud
10. Mobile BI and analytics
11. Analytics platforms

The trends possess a vast amount of benefits that are potentially worth of the effort to be taken. These trends not only save money and time but also make the analyst job more interesting because now there would be more time to actually analyze the data instead of collecting it.

1.1 Background and Motivation

My background on this topic is related to my work experience in the financial sector and bachelor's thesis subject on big data. While researching big data in my bachelor's thesis it was a natural continuum to research the companies' maturity level and the available solutions for them in this master's thesis. I have been working in the financial sector since

2011 and as an information and service management student I know that data is currently one of the key assets in every company. The positions are more and more focused on data analysis. There are many types of data such as transactional data, customer data, financial data and external third party data. Managing these kinds of large data sets becomes more complex every year, because the amount of data increases exponentially every year. Even if there is more advanced software available that helps us to handle the data, it requires professionals to constantly educate themselves to keep up with the development.

My motivation is high for several reasons. Firstly, I am interested in business intelligence and analytics development and I have been studying these topics intensively during my master's degree studies. Secondly the senior management team of the case study company has identified five top business challenges in the company to focus on towards 2016. One of these challenges is labelled as "fact based decisions". In short the idea is that decision making processes should be more based on "facts" and data and less based on common sense. Since the executive level has proved their commitment it naturally increases the overall motivation of the development team. Thirdly, I feel like building modern BI&A solutions in 2015 is very ambitious. Building BI&A solutions might be complex and time consuming but at the same time rewarding once it is finished. The solutions help modern society to be smoother than before in addition to increased brain capacity.

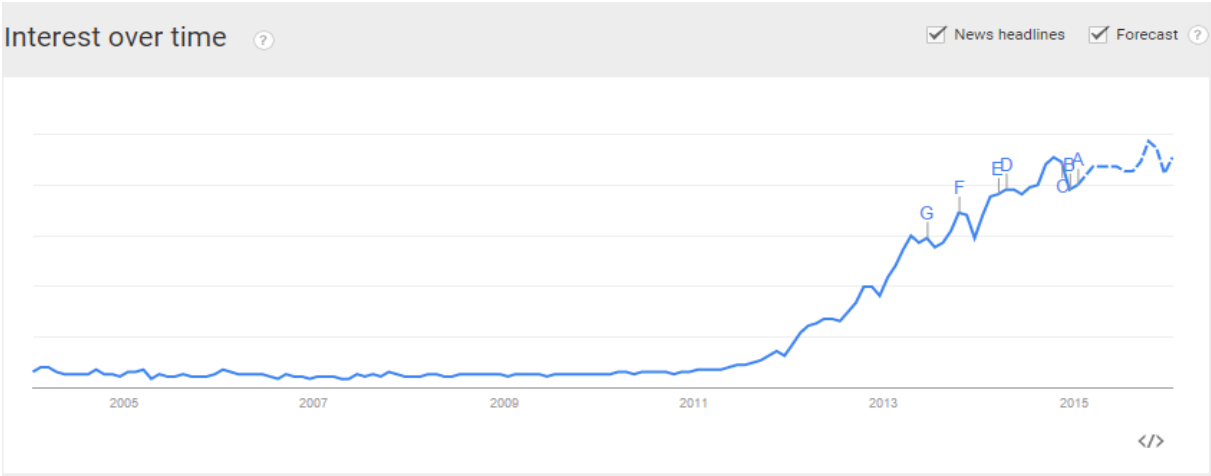
In many cases the development of business intelligence and analytics could potentially enable automating a large number of manual tasks. At this point there is a SAS and Excel combo to create the reports in the case company and there is a lack of automation as well. The problem is that Excel cannot efficiently handle large amount of data instead it is slow and vulnerable to crashes. My goal is to conduct a future strategy regarding the BI&A. I will find out where the company is standing at the moment. Then, I will investigate how to move on from current state of analytics. Researching this topic will be interesting and I am keen on to read what is the next top notch solution in future from the academic perspective.

My motivation to do a research about BI&A and modern data warehousing is also related to big data because it is part of the future concept in many industries. I am enthusiastic to understand our current business environment and how the managers are making their business decisions. Information is very closely related to the decision making and it provides good support for the managers. Even though information has always been used, recently we have realized that the amount of data is massive and the amount is growing

all the time at a fast pace. In order to keep up with this technical and data generating development it is desirable to study the topic more closely.

For companies, Modern BI&A solutions and big data could potentially be the next way to stay on the edge of peak performance due to more advanced ways of using data. In addition, according to Google Trends big data (figure 1-1) is at its all-time high in hits. The popularity shows also in the amount of big data articles. For example, there are many articles available that have been published between the years 2011 and 2014 and the trend is likely to continue this year also. These are the reasons which make big data so interesting.

Figure 1-1 – Big data search interest over time (Google 2015)



1.2 What is Business Intelligence and Analytics?

Analytics requires the ability to collect, manage, analyze, and act on ever-increasing amounts of disparate data, at the right speed and within the right time frame. It includes methodologies for development as well as technologies. When people think about analytics, they generally consider a range of techniques, including spreadsheets, query and reporting, dashboards, performance management, and more advanced techniques such as visualization and predictive modeling. Analytics can be divided into two parts: Business Intelligence and Advanced Analytics (Halper and Stodder 2014, 5.)

Business Intelligence could be considered as historically oriented transactional, financial, profit/loss, and cost-management activities. Because visual presentations are part of BI, data visualization, user interfaces, and the user experience with data generally on workstations, laptops, and mobile devices are part of BI. BI's definition often extends to the server. A BI server could include OLAP cube creation/management, ETL, and other data warehouse functions. Self-service, visual data discovery technologies are changing the face

of BI by enabling users to do more with dashboards, reporting, data analysis, and visualization on their own with less IT direction (Halper and Stodder 2014, 5.)

Advanced analytics provides algorithms for complex analysis of structured or unstructured data. It uses sophisticated statistical models and formulas, machine learning, and other advanced techniques to find patterns in data for prediction and decision optimization. As analytics becomes more advanced, it often becomes more algorithmic. Of course, analytics is not just about techniques. It includes the infrastructure and data management to support disparate kinds of data from a variety of internal and external sources. It also includes the cultural and organizational processes that enable companies to become more data driven. This includes development techniques as well as the processes in place to manage, govern, and utilize the data and analysis by a wide range of people in the organization (Halper and Stodder 2014, 6.)

1.3 What is Big Data?

Big data is a rather new term which is not commonly used outside the field of Information technology. The closest synonym for big data is data or business analytics. However, big data can be separated from the concept of analytics by three key differences: High Volume, High Variety and High Velocity (McAfee & Brynjolfsson, 2012). According to Jukka Ruponen (2012), several companies have added a fourth “V” that they have coined differently: Gartner named it "Virtual" to include only online assets in big data. IBM named it "Veracity" to identify highly varied accuracy of the data. Finally, Oracle named the fourth V as "Value" to identify the challenge in turning big data into economic value. “Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze” (MGI 2011, 2).

1.3.1 Volume

Thomas Davenport (2012, 43) claims that there is no doubt about it that organizations are swimming in an expanding sea of data that is too voluminous. As mentioned earlier in this introduction part, it is hard to realize how much data is out there. In 2012 altogether 2.5 exabytes of data were created each day, which equals with 25 million library floors of academic journals. Of course not all of this data is useful for every company but we need to be aware of how enormous our playground is.

Basically, big data can be divided into two parts. An information ocean is a place where huge amounts of data are stored and analyzed later. For example data warehouses can be described as information oceans. However there are also information streams, for example social media, click streams, sensor data, and emails. Streams are not always stored; instead these streams may be analyzed real-time and after that the data dismisses. Even though the volumes are huge, we do not have to store it all. (Ruponen 2012.)

1.3.2 Velocity

For many applications, the speed dominates the volume (McAfee & Brynjolfsson 2012, 63). It is important to have real-time information to be able to make more accurate and agile decisions. Information streams mean that information is moving in high speed in other words in real time. One big challenge is to catch the data in real time and refine it into knowledge (Chen et al. 2012, 1167). That is great opportunity for businesses that spend enormous amounts of time and money to try and understand their customers (Sheridan 2009). The faster you receive information the faster you can react for example to market demands.

1.3.3 Variety

One characteristic of big data is its unstructured forms. According to Laney (2011) there is no greater barrier to effective data management than the variety of incompatible data formats, nonaligned data structures, and inconsistent data semantics. Big data covers such forms as: web site clicks, updates, images posted, GPS signals, online-shopping, sensors, instruments, real-time logs, money trades, other high speed transactions, likes, profile updates, opinions in online forums, all social media contents, blogs, tweets and many more. All this in addition to traditional standard data types can be included to big data (Davenport 2012, 43-44; Ruponen 2012, 6-11).

Big data has been also criticized as if the term itself includes too much of everything. According to Immo Salo (2013) one of the problems is that there is no real consensus regarding the big data. People may talk different language with each other even though the topic is the same. This can cause misunderstandings and it can create confusion in the field. As McAfee and Brynjolfsson (2012, 63) put it “Each of us is now a walking data generator”, this quote sums up nicely where all of this data is bursting from, all the devices we carry with us such as smart phones and iPADs are creating data to external parties. Mobility of digital data has a significant impact to the amount of big data.

All of the characteristics above make the field of big data look like a data jungle. There is a lot to digest for companies that want to make the most out of it. But the premise is same for every company: everyone gets the input (big data) and between input and output every company has its own processes. The better the firms are able to refine this input, the better and finer the output is. At the end that is all that matters. The refining process of big data is the future key to gaining competitive advantage.

1.4 What is a Data Warehouse?

A data warehouse (DW) is a pool of data produced to support analytical and managerial decision making; it is also storage of data from present and past, data of potential interest to managers throughout the organization. Data are usually structured to be available in a form ready for analytical processing activities such as online analytical processing (OLAP), data mining, querying, reporting and other decision support applications. A data warehouse is a structured, integrated, time-variant and stable collection of data in support of management's decision making process (Turban, Sharda & Delen 2011, 329.)

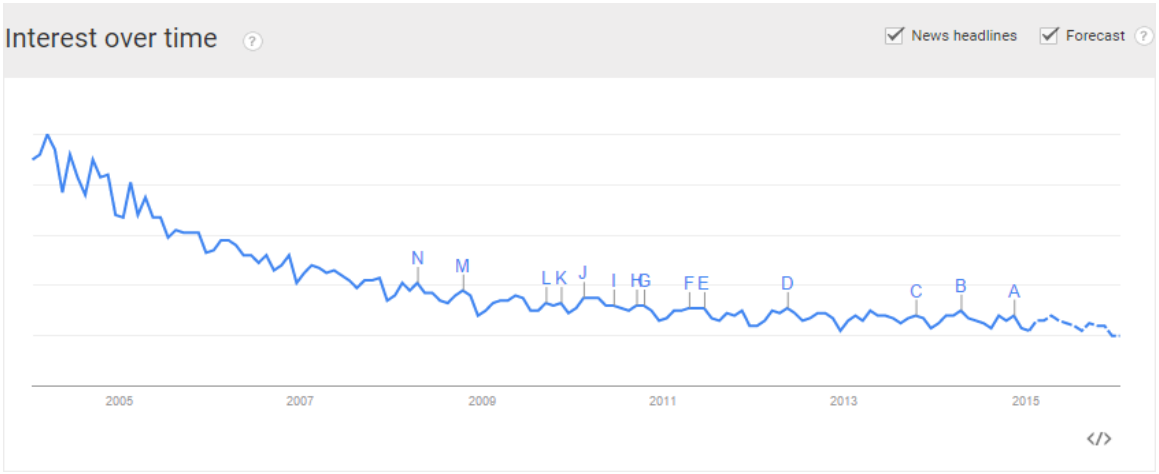
As Williams (2014) puts it "DW is a specialized database used to store important business information about transactions, products, customers, channels, financial results, performance metrics, and other business information over multiple years so the data can be easily and consistently used to improve business results".

Data warehouses can be categorized based on its limitations whether it is department or enterprise wide. For example a data mart (DM) is usually a smaller version of DW and it focuses only in single subject area (e.g., risk, finance or sales). The DM can be either dependent or independent. A dependent DM (DDM) is directly linked to a data warehouse whereas independent DM (IDM) gets the data from elsewhere. IDM can be described as scaled down version with lower costs. The DDM ensures 'single version of truth' within a company and it also provides consistent data model and high quality data (Turban, Sharda & Delen 2011, 330.)

Enterprise data warehouse (EDW) is a company wide solution and it contains data from all departments. Although there are many interpretations of what makes an enterprise-class data warehouse, the following features are often included: A unified approach for organizing and representing data, the ability to classify data according to subject and give access according to those divisions e.g. sales, finance, inventory and so on (Janssen 2015.) The main difference between DM and EDW is that EDW gives leaders a fresh, big-picture

perspective with a 360-degree view of the business whereas DM focuses only one part of the business.

Figure 1-2 – The Data Warehouse search interest over time (Google 2015)



In contrast to big data interest, DW interest has been decreasing many years (see Figure 1-2). According to D'Antoni & Lopez (2014) the concept of data warehousing has reached maturity within IT, companies and among data analysts. Mainly because DW has been discovered almost 30 years ago it is not a surprise that companies have been able to familiarize themselves with the concept. Sarsfield (2009, 10) criticizes data warehouse systems for various reasons: 1. Require more resources, both in technology and expertise, to repeatedly extract big data sets and work out disparities between them. 2. Source data remains disparate and managed by the rules of the individual business unit. 3. Does not solve the problem of having to separately manage data silos with additional people. 4. No centralized process for improving data enterprise-wide. 5. Business Intelligence is rarely real-time.

On the other hand according to D'Antoni & Lopez (2014, 11) "pundits have speculated that big data platforms could be the death of the traditional data warehouse". However, from my own experience there are still many financial, regulatory and ad hoc reporting requirements that will ensure that the data warehouse remains a component of the IT landscape in the future. Even though this is rather well-established concept, for these reasons, companies are still finding it worth to use. However, during the last few years the concept of big data and the technologies developed around it such as Hadoop, require new, modern data warehouse architecture. This new wave makes data warehousing still a very relevant topic that we shall not bury quite yet.

1.5 What is Data Modelling?

Data model is a representation of a real world situation about which information is to be collected and how it will be stored by applying formal data modelling techniques. Data modelling involves professional data modelers working closely with business stakeholders, as we as potential users of the information systems. Critical part of data modelling is to define and analyze business requirements that are needed to support the business processes (Helenius 2014, 12.) Data modelling is very time consuming and it should not be done in a rush. From my own experience there is usually a lot of thinking and testing involved whether the data model is working appropriately or not. Most importantly the data model has to meet the customer and user needs as well as the technical solution.

1.6 Objective of the Thesis

The main objective of this thesis is to provide an answer to the following research questions:

1. *What does the state of art modern BI&A solutions look like today?*
2. *What is the current BI&A maturity level of the case company?*
3. *What kind of future BI&A-strategy would be optimal based on the literature review, survey results and architecture analysis?*

This study goes through scientific articles, current ICT debates and professional opinions accompanied with case study and maturity test and picks up elements that are significantly affecting the data warehouse implementation process. This thesis is also providing an alternative view on what kind of architecture data warehouse could have in near future. What are the differences between classic and modern architecture? How can we combine the current big data hype and already mature data warehouse scene? There is a minor research gap on measuring the maturity level of BI&A in Finnish companies and there is no basic research done on how to implement modern BI&A solutions. My purpose is to provide more knowledge on the foremost topic.

Due to the limited length, focus and scope of this study, this study is not going to provide any actual tool descriptions or functions; instead it is focusing on the process itself and hopefully the study will point out new questions to be answered. This study aims to address the challenges with the developed model and also aims to invite reflections regarding the field of data warehousing and big data.

1.7 Methodology

This thesis consists of two parts, a literature review and a case study which includes maturity evaluation and a proposal of a future strategy. The weight is distributed evenly on both parts. In the literature review section the study aims to find the most common ways to model the data and implement the classic and modern DW system. Studying the process in a real life is likely to bring the research insights and sense of practicality to this thesis. The professional opinions from consultancy and management level aim to bring real-life experiences and credibility to the proposed strategy.

The theoretical framework is based on the articles related to the architecture, best practices and future alternative methods. Professional opinions and real life cases are commonly introduced in these articles. There are no quantitative models used in this study. Instead the quantitative models are more related to the mathematical tools, predictive and probability analysis around this topic. Therefore the theoretical framework is aimed to be strongly qualitative. The framework is approached from three different angles: current way of doing things, alternative ways and future best practices. I think these three different points of view give this thesis an appropriate structure that helps the reader to perceive the overall picture.

The strategy is planned for an anonymous company that is currently developing its business intelligence and analytics system. I aim to study the current architecture, implementation process, and best practices. Then I am suggesting a future strategy on how to move on. I am part of the development team and therefore I have personal biases towards this study and strategy. I aim to bring those biases above surface in order to keep this study neutral enough. The use of deductive qualitative approach in this research that means that the order of the study is following: New Knowledge > Proposal of Usage > Evaluation. However the evaluation of usage cannot be done due to limited amount of time. New knowledge in this case would be the information bursting from scientific articles.

1.8 Structure of the Study

The first chapter introduced the background and motivation, thesis objective, methodology and the concept of analytics, business intelligence and advanced analytics, big data, data warehousing and data modelling. The second chapter presents the theoretical framework of my topic. It includes literature review that includes: fact-based decision

making, data governance, classic data warehouse architecture, data modelling methodologies for data warehousing, and analytics in data warehouse environment. Then the second chapter moves on from data warehouse implementation and development to modern data warehouse architecture and finally to advanced analytics in modern data warehouse architecture environment. The third chapter goes through the methodology in great details. Fourth chapter reviews the case study and its results, including the proposed future strategy. In the final chapter the thesis is summarized and conclusions are made. I hope that the dialogue between theory and practice can be observed by the reader of this research.

2 LITERATURE REVIEW

This chapter explains the essential concept of data governance that is required for fact-based decision making. It is important to understand how we should manage the data in 2015 in order to build advanced information systems efficiently for the needs of decision making. After the introduction to data governance, the literature regarding the data warehouse architecture, data modelling and implementation process will be gone through. These three sections are the most important and time consuming parts. In order to successfully maintain the data warehouse environment and the business around it, one should focus on these areas, and that is why those are in the center of this thesis' theoretical framework. Finally, the advanced analytics methods in modern data warehouse environment will be introduced.

2.1 Fact-Based Decision Making

Executives make tough decisions every day with inaccurate information and limited resources under pressure. Too often these calls are made by analyzing irrelevant and unreliable data. Organizational politics, formal authorities, or plain plausibility may lead companies to choose the wrong path. The loudest character or the highest-paid person's opinion weights the most and decisions are made based on gut feel. Unfortunately, this intuitive and often personality-driven approach is the norm in large corporations all over the world in Europe, the US and Asia (Kelley, 2009).

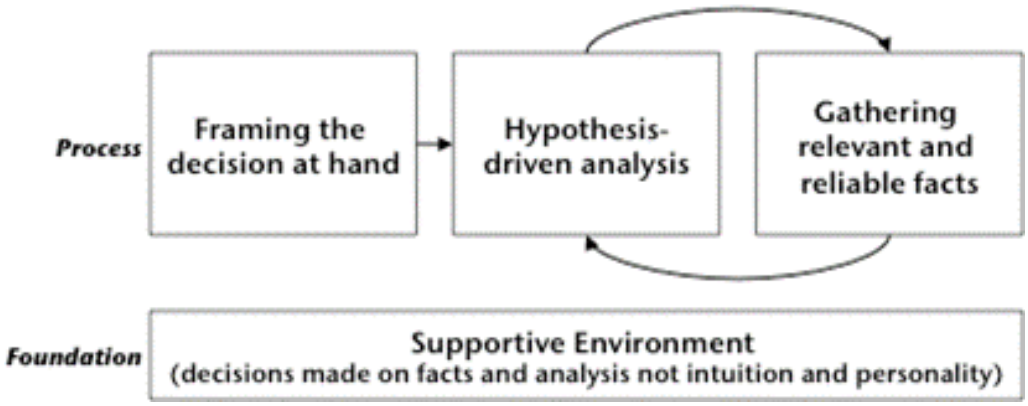
There is an alternative approach called fact-based decision making (illustrated in Figure 2-1), also known as evidence-based management or data-driven decision making. "Fact-based decision making is both a methodology for executive decision making under uncertainty as well as a philosophy of how to tackle business problems" (Kelley, 2009). There are four key characteristics identified by Kelley (ibid.):

1. The decisions are based on facts and analysis - not intuition and personality.
2. The structure of the decisions is well understood, and the decision is carefully framed to reflect the reality of the outside world.
3. Analysis is hypothesis-driven, focused in areas that drive choices between options.
4. Data gathered to support the decision is relevant and reliable and preferably real time.

. Fact-based decision making has been proved as a valid approach and option for companies. McAfee and Brynjolfsson's (2012, 64) research at MIT Center for digital

business show that companies making data-driven decisions are 5% more productive and yield on average 6% more in profits.

Figure 2-1 – Key elements of Fact-Based Decision Making (Kelley, 2009)



To determine the profitability and transparency of a new product, for example, relies on knowing where we are making money, where not, why, who will buy the product, how much they will be willing to pay and where are the opportunities to improve. Already pricing decisions alone require in-depth knowledge of demand elasticity, costs, and customer economics. Fact-based decision making is the theoretical approach applied to business decisions.

McAfee and Brynjolfsson (2012, 66) argue that even though data-driven approach may give a company a great competitive advantage it does not remove the need for vision or human insight; instead the basic leadership skills such as spotting a great opportunity, persuading people to work hard and thinking creatively are still needed despite the fact that organizations move towards data-driven decision making. The same conclusions were made out of the managerial interviews related to big data (Kulin 2013, 16), for instance one of the executives note that people management is done by intuition and the actual business related decisions are based on pure data.

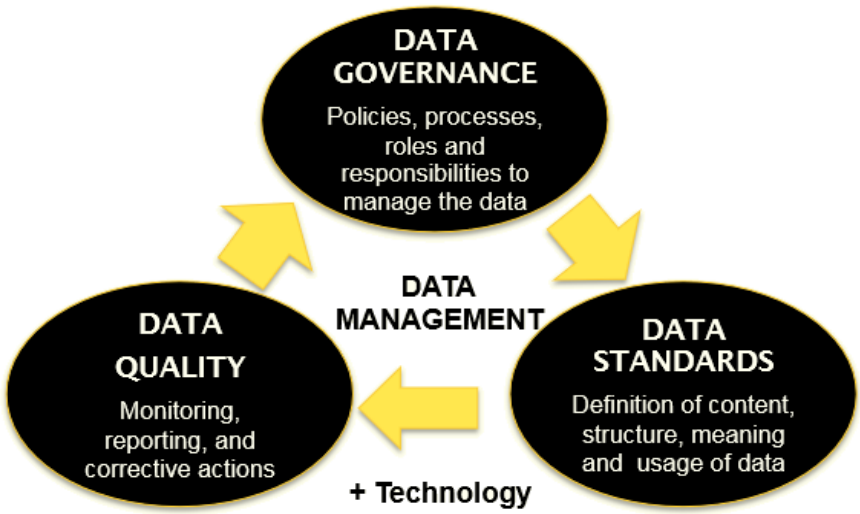
Fact-based decision making should enable one version of the truth. Thus, the companies would be able to avoid the controversy related to data. Secondly fact-based decision making improves the odds of making a good call that would lead to a successful outcome because the decision is based on pure and single data. After acknowledging the benefits of the approach, companies need the actual, correct data. The next section is

explaining the basics of data governance, which is a rather new term. Data governance has evolved from the data management synonym that has been more commonly used in literature.

2.2 Data Governance

If fact-based decision making is defined as part of the organization culture like e.g. innovation there is the same need for data governance committee as if there is a need for innovation committee. In practice data governance is steering the actions related to data management from top and it oversees the entire journey of data. Niemi (2014, 18.) describes data governance as a collection of best data management practices that orchestrates business and IT to work together in order to ensure the uniformity, accuracy, stewardship, consistency and accountability of the enterprise's core data assets. Based on his experience approximately 80% of it is about people and processes and 20% about technology. Data governance provides formalized discipline to ensure accountability for the management of company's core information and provides structure and sponsorship for decision making (Niemi 2014, 18). See figure 2-2 for a rough overview of the data governance.

Figure 2-2 – Data Governance Flow (Niemi, 2014)



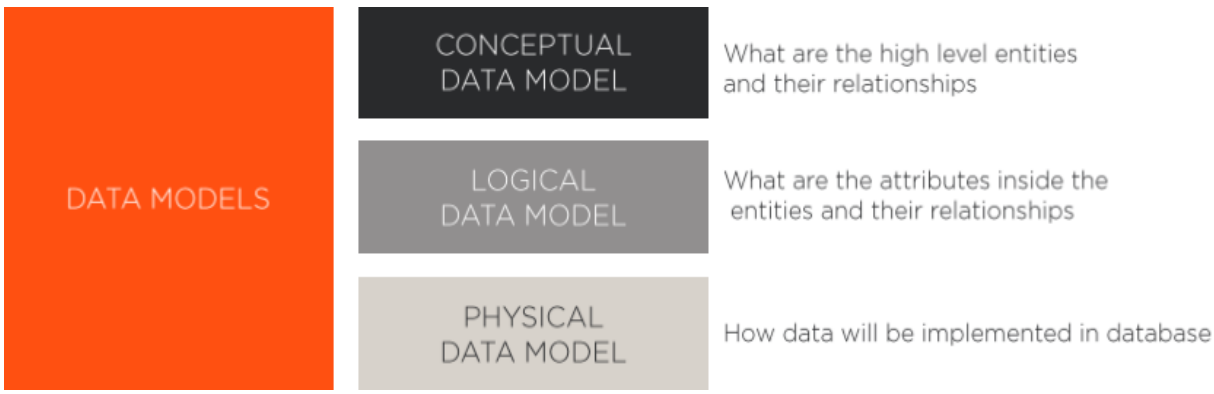
2.2.1 Data Standards

As we can see from the figure 4, under the data standards are content, structure, meaning and usage of data. These components are the bones, muscles and organs that keep the body of data together. Further, one could describe the enterprise data model (EDM) as DNA or brains that determine how well the muscles, bones and organs are functioning as a unit. Finally, the Meta data modeler that creates and maintains the organizational data model

could be described as God. According to Kendle (2005) an EDM is an integrated view of the data produced and consumed across the organization. It "represents a single integrated definition of data, unbiased of any system or application. It is independent of 'how' the data is physically sourced, stored, processed or accessed. The model unites, formalizes and represents the things important to an organization, as well as the rules governing them". An EDM is data architecture, a framework that is used for integration. Integrated data is meant to provide a "single version of the truth" for the benefit of all. It minimizes data redundancy, disparity, and errors; core to data quality, consistency, and accuracy (Kendle 2005).

Data models can be divided into three categories (see figure 2-3, Helenius 2014): conceptual, logical and physical data model. The models can be visualized like an architectural blueprint is to a building. These three model types are illustrated in more detail and separately in appendix A. In short the simplest level is the conceptual one and the complexity is increasing while descending towards the physical data model. Usually conceptual level does not include any technical names so that executives and leaders at all levels can understand the fundamentals of the data body. On the other side of the coin is the physical model which is very detailed in technical terms and is useful to both IT-architects for design means and data miners that use for example standard query language (SQL). Logical data model is useful for semi data-oriented people who are for example involved in process development. They do not have to know the detailed technical aspects, but are required to know what data is available and what logic is behind it.

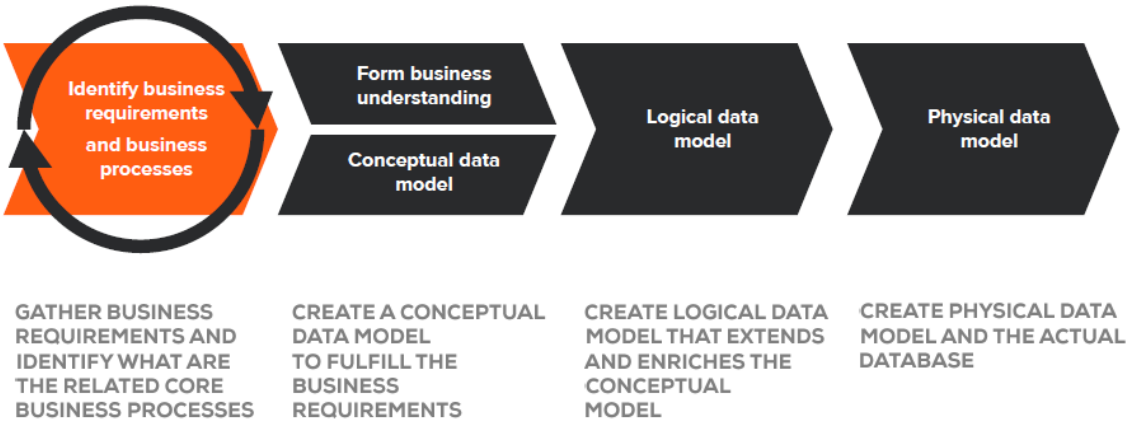
Figure 2-3 – Data Model breakdown by levels (Helenius 2014)



As a framework for data architecture, an EDM is the starting point for all data system designs. "For enterprise data initiatives, such as an operational data store (ODS) or data warehouse, an EDM is mandatory, since data integration is the fundamental principle underlying any such effort. An EDM facilitates the integration of data, diminishing the data

silos, inherent in legacy systems. It also plays a vital role in several other enterprise type initiatives" (Kendle 2005.) In figure 2-4 we can see the data modelling development cycle (Helenius 2014). It begins with business requirements and ends up to the very detailed physical data model. Data modelling process will be discussed in-depth under the chapter 2.4.

Figure 2-4 – Data Modelling development cycle (Helenius 2014)



2.2.2 Data Quality

Once the data model is created and tested it is ready for use. During the operational usage period data governance maintenance tasks include monitoring, reporting and corrective actions in order to improve the quality and usability of data. There are several KPIs identified for measuring and ensuring the quality of data. Accuracy: correctly reflects the real world object. Completeness: expected attributes are provided. Consistency: in sync across the enterprise. Coverage: covers expected amount of data. Uniqueness: no duplicates within or across systems. Timeliness: data delayed is data denied. Auditability: can be tracked to originating transactions. Stressing the quality of data is not exaggerated because it has far reaching effects when decisions are made based on pure data. When it comes to ensuring the one version of truth, quality is the key.

The quality feedback from users and system should be examined in the steering committee of data governance function. Once discussed the cycle shown in figure 2-4 continues and should be constantly iterated. The data model can be modified and updated. One thing that is not mentioned in Niemi's (2014) theory is training. As it is some sort of norm that tools are developed almost every year, training is essential in order to have high quality of data. When employees know how to use the advanced tools correctly and

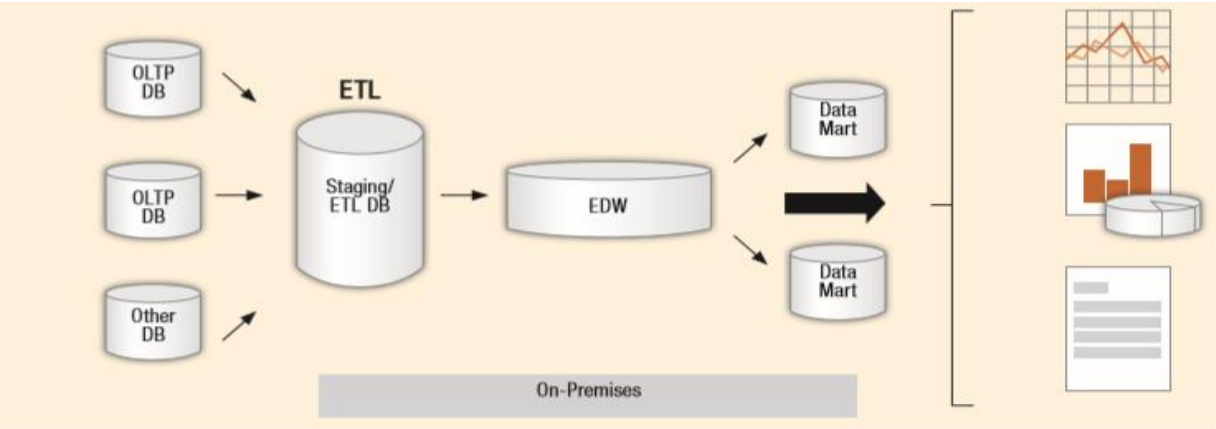
efficiently it improves also the quality of data that is used to make decisions. There might be even new methods to use or analyze the data like in case of big data.

2.3 Classic Data Warehousing Architecture

The data warehouse has become one of the most widely used applications of advanced database technology today and it is recognized as one of the most efficient methods to process data (Kaldeich & Oliveira e Sá 2004, 2). The new era of company-wide systems integration and the growing demand towards business intelligence both accelerate the applications. Many of the large companies have established data warehouse systems as a component of their information systems landscape (Guo et al. 2006, 59). In Figure 2-5 D'Antoni and Lopez (2014, 12) illustrate the high level, classic data warehouse architecture. The architecture shows that there are five different phases in the DW process flow.

In the first phase, the data are located in the databases in its original form. This can be any data related to the business for example online transaction processing (OLTP), customer data, and enterprise resource planning (ERP) or web data. The data in different locations are usually structured in custom way.

Figure 2-5 – Classic Data Warehouse Architecture (D´Antoni and Lopez 2014)



From the databases the data is moved, or in other words extracted, to the staging area. In this second phase a time consuming ETL process (extract, transform and load) is required. According to Turban, Sharda & Delen (2011, 334) data are extracted using custom-written or commercial software. On a quick view there is a myriad of commercial software available on the market. The top 15 list (coined by predictiveanalyticstoday.com, 2014) includes such software as Jaspersoft ETL, Talend Open Studio, HPCC Systems and many more. This is part of the phase two known as ETL.

In the staging area the data is cleansed and transformed according to certain set of rules or functions. After this the data should be ready to load into the end target, usually the data warehouse. In short, ETL refers to the process of extracting data from databases or outside sources, cleaning and transforming it, and finally loading it into the end target database, more specifically, the data warehouse. The ETL may take up to 70% of the whole DW development time. The most crucial part before the load is data modelling that includes fact table and dimensional table models. A few data modelling theories and methods are introduced under chapter 2.4. Below is only a short description of the usual design and objectives.

According to Moody and Kortink (2000, 3) "The objective of dimensional modelling is to produce database structures that are easy for end users to understand and write queries against. A secondary objective is to maximize the efficiency of queries. It achieves these objectives primarily by minimizing the number of tables and relationships between them. This reduces the complexity of the database and minimizes the number of joins required in user queries". See figures 2-6 and 2-7 for the most common dimensional model called star schema. Oracle (2000, 16) claims that "The star schema is the simplest data warehouse schema. It is called a star schema because the diagram of a star schema resembles a star, with points radiating from a center. The center of the star consists of one or more fact tables and the points of the star are the dimension tables". Dimension tables provide the basis for aggregating the measurements in the fact table. The fact table is linked to all the dimension tables by one-to-many relationships.

Figure 2-6 – Star Schema model example 1

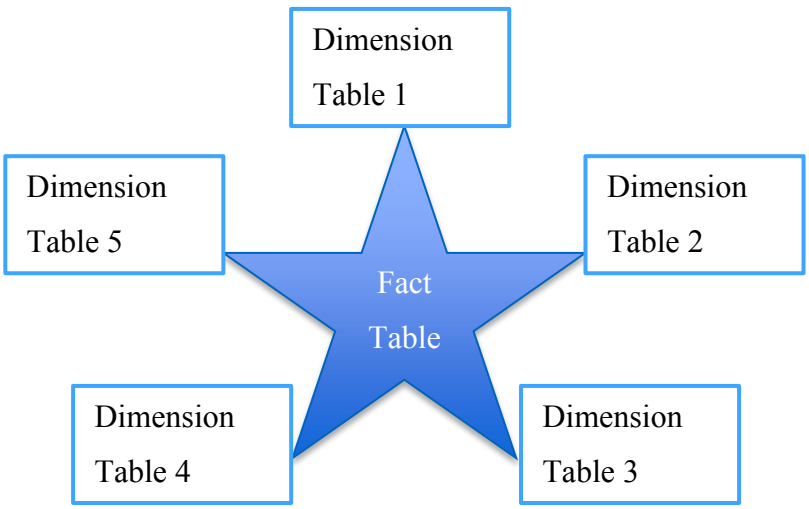
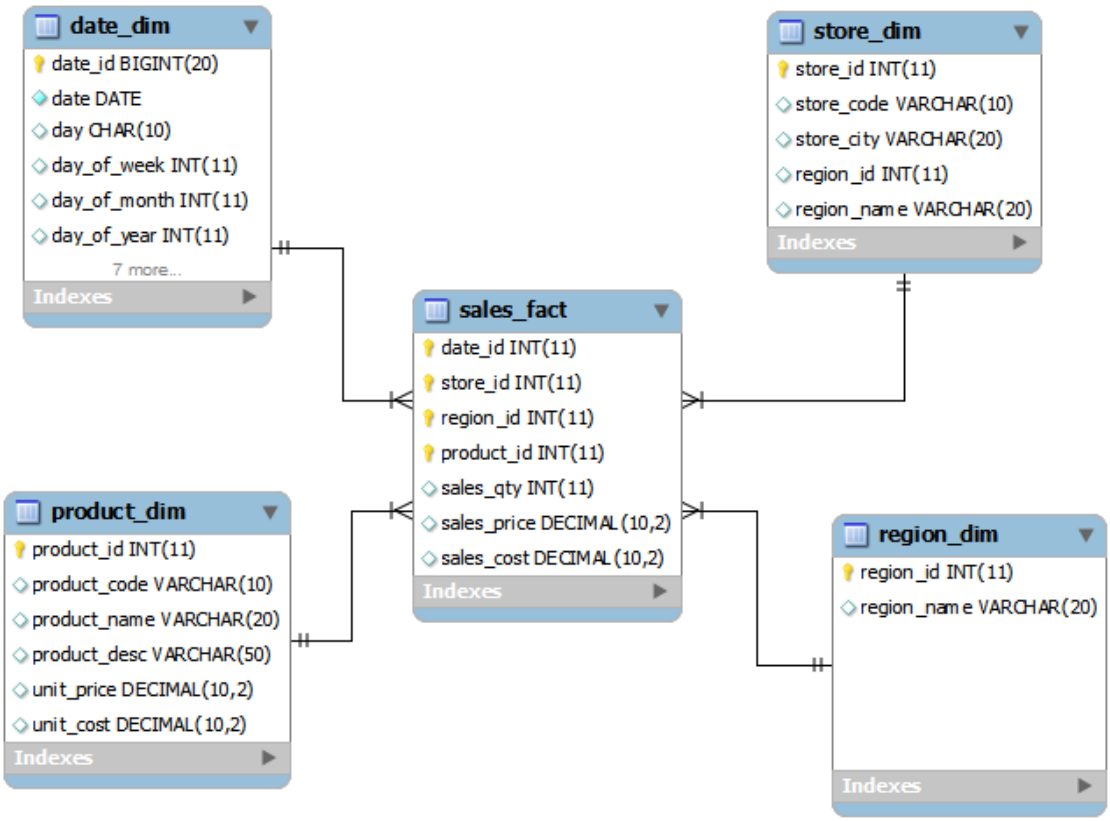
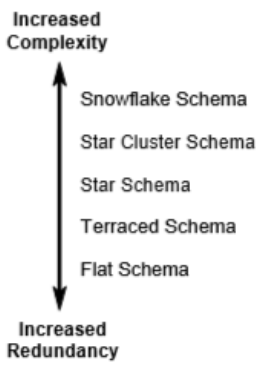


Figure 2-7 – Star Schema model example 2 (DWHworld.com 2010)



Star schemas may either be implemented in specialist OLAP (Online Analytical Processing, computer-based technique for analyzing business data in the search for business intelligence.) tools, or using traditional relational DBMS (Database Management System, system software for creating and managing databases). Accompanied by Turban et al. (2011, 351) "Many variations on the data warehouse architecture are possible; the star schema is the most important". Moody and Kortink (2000, 11) agree on Turban's claim that there is a wide range of options for producing dimensional models. These include: flat schema, terraced schema, star schema, snowflake schema and star cluster schema. Each of these options represent different trade-offs between complexity and redundancy (see figure 2-8).

Figure 2-8 – Design Schemas' Tradeoffs (Moody and Kortink 2000)



Besides the actual data modelling, the data modelling process includes a lot more. Once the first draft of data model is finished, it needs to be tested. When the data is loaded into the EDW, reconciled and tested in production environment, the hardest work of development has been done. The development work is followed by maintenance. Maintenance usually includes such actions as updates to the models, taking care of servers. In other words users may want to include new products in the model or some changes needs to be made to the original design. A new round of testing is needed after every change that is made. Therefore maintaining DW is an ongoing process which requires resources not only during the development work but also after it. However this is the case only if the original model is near flawless. Yaddow (2014, 30) stresses the importance of getting the data correct because fixing it takes a great deal of work. His team claims to have seen models so corrupt that it is better to start from scratch. According to Jukka Ruponen (2014) one of the concerns that DW produces is that the "meta data modeler becomes now the most important man in the house". It is now even more crucial for the company to consider what-if scenarios for this particular person if he is about to leave the company. Deep understanding of company data and its architecture is hard to substitute and time consuming to gain.

2.3.1 Data Marts

Companies can consider restricting the usage of data according to its user base via data marts. This arrangement can be done to create custom data for division specific needs or for example for security reasons. As Moody and Kortink (2000, 4) puts it "Data marts represent the 'retail' level of the data warehouse, where data is accessed directly by end users. Data is extracted from the central data warehouse into data marts to support particular analysis requirements. The most important requirement at this level is that data is structured in a way that is easy for users to understand and use. For this reason, dimensional modelling techniques are most appropriate at this level. This ensures that data structures are as simple as possible in order to simplify user queries".

Saunders (2009, 20) is on same page with Moody and Kortink (2000, 4) "in our kitchen, a data mart would be like food that is partially pre-made to expedite completion of the dish". Data is collected from the central data warehouse into data marts to support specific and particular analysis requirements. The most important requirement at this level is that data is structured in a way that is easy for users to understand and use. For this reason, dimensional modelling techniques are most appropriate at this level. This ensures that data

structures are as simple as possible in order to simplify user queries (Moody and Kortink 2000, 4.)

Using data marts is faster also from division's perspective, because you do not have to process full data every time instead the division can choose their own fact table setups and move on from there. From security point of view it might be a good idea to not share everything with everyone. As stated in one of the managerial interviews regarding big data (Kulin 2013, 17), one great risk for companies are the moles that are using the enterprise information in harmful ways.

When the final call for ready-to-use-data has been made, users such as analysts and managers can begin to use the data through the analytical tools. For users the analyzing should now be faster, simpler and easier than before. The software add-ons built on data warehouses to support it like Qlikview, IBM Cognos and SAP are widely used in companies and this is the current trend that is pushing companies away from using the Excel spreadsheets. In future the big data analytics are likely to change the classic data warehouse architecture and built on applications are going to transform the analysts' methods to answer the business questions. To name a few Hadoop, InfoSphere, SAP Hana, Splunk, Pentaho Business Analytics and JasperSoft are only the tip of an iceberg in the developing big data market. The big data impact will be discussed later on under sections 2.7 and 2.8.

2.4 Data Modelling Methodologies for Data Warehousing

One of the greatest challenges in data warehousing is how to develop flawless data models to support complex querying, reporting and analysis. Although great achievements in research have been achieved on data warehousing, there is still a need for more techniques such as active rules, update filtering, parallel processing, data expiry, data indexing and some other items that would help to accomplish the flawless model Nguyen et al. (2005, 532). In practice also, data modelling is a very demanding task and requires a lot of co-operation and brainstorming among modelers and DW end users. This all is related to the data modelling; therefore, it is still useful to research data modelling methodologies in DW.

2.4.1 Kimball's Dimensional Data Modelling for Data Warehouse (1996)

Ralph Kimball from the Stanford University, who came up with the very first dimensional models in the late 1990's can be called the father of data modelling. Traditionally until the early 21st century the DW development has been categorized to three

different approaches: data-driven, goal-driven and user-driven each as a single approach with no link to one or another (Kaldeich & Oliveira e Sá. 2004, 3). According to Kimball (1996, 1997), the data warehousing (OLAP) environment is profoundly different from the operational (OLTP) environment and techniques used to design operational databases are inappropriate for designing data warehouses. Given this premise, Kimball proposed a new approach for data modelling specifically for designing data warehouses, called dimensional modelling. The method was developed based on observations of practice, and in particular, of data vendors who are in the business of providing data in "user-friendly" form to their customers. The method has never been empirically tested, but has clearly been very successful in practice (Moody & Kortink 2000, 2). Dimensional modelling has been adopted as one of the most significant approach to designing data warehouses and data marts in practice. It also represents an important contribution to the data modelling and database design.

Data-driven data modelling in DW starts with analyzing the transactional data from a source database in order to reengineer their logical data schemas. Data-driven data modelling constructs DW data models based on operational system database schemas while overlooking business targets and user requirements. This raises two questions: 1. How to analyze transactional data sources and match them with information requirements and data warehouse data models? 2. How to reorganize the identified source schema elements to form data warehouse data models? Some studies suggest that dimensional models such as Star and Snowflake are used to reorganize data source schemas (Guo et al. 2006, 59.)

Goal-driven approach places emphasis on the need to align data warehouse with corporate strategy and business objectives. Goal-driven data modelling forms data models based on business goals and accorded business processes ignoring data sources and user needs (Guo et al. (2006, 60.) Rob Weir (2003) claims that nineteen articles that referred to DW implementations before 2000, fifteen of them communicated that the implementation should meet and agree on corporate strategy and business objectives. Only few if any article try to answer how to engage business strategy and business targets to data warehouse data models.

User-driven approach requires involvement of end users in data warehousing. Basically data modelling derives data models directly from user query requirements without considering data sources and business goals. The problem is that this approach is not focusing on how to transform user needs into appropriate design elements (Guo et al. 2006, 60.)

These approaches are aimed to determine information requirements of data warehouse users. End users alone are able to define the business goals of the data warehouse systems correctly. Therefore end users should be enabled to specify information requirements by themselves. However, end users are not capable of modelling the Meta data because they cannot have sufficient knowledge of all available information sources, and because they use only a business unit specific interpretation of data (Kaldeich & Oliveira e Sá. 2004, 4.) As described above all approaches have positive aspects and all of them raise challenging questions, and therefore they are not perfect alone. A few studies' objective has been to merge all positive aspects to a new approach. In the section 2.4.3 I will go through one of them in more detail.

2.4.2 Data Warehouse and Data Mart Design by Moody and Kortink (2000)

Moody and Kortink (2000, 11) describe their method for developing data warehouse and data mart designs from an enterprise data model as a state of art solution that has been applied in a wide range of industries, including manufacturing, health, insurance and banking. They claim that "Entity relationship modelling is equally applicable in data warehousing context as in an operational context and provides a useful basis for designing both data warehouses and data marts". Moody and Kortink's model is supporting the classic data warehouse architecture.

The Moody and Kortink's method has evolved considerably as a result of experiences in practice. Moody and Kortink (2000, 4) argue that "different design principles should be used for designing the central data warehouse and data marts and for example central data warehouse design represents the "wholesale" level of the data warehouse, which is used to supply data marts with data". On the other hand data marts represent the "retail" level. The most crucial requirement of the central data warehouse is that it provides a consistent, integrated and flexible source of data. The traditional data modelling techniques (entity relationship models and normalization) are the most appropriate at this level (Moody and Kortink 2000, 4).

The sparkle behind their model is based on a set of challenges. One of the challenges with using traditional database design techniques in a data warehousing environment is that it results in database structures that are too complex for end users to understand and use. A traditional operational database consists of hundreds of tables that are linked by a complex web of relationships. Even simple queries will require multiple table joins, which are

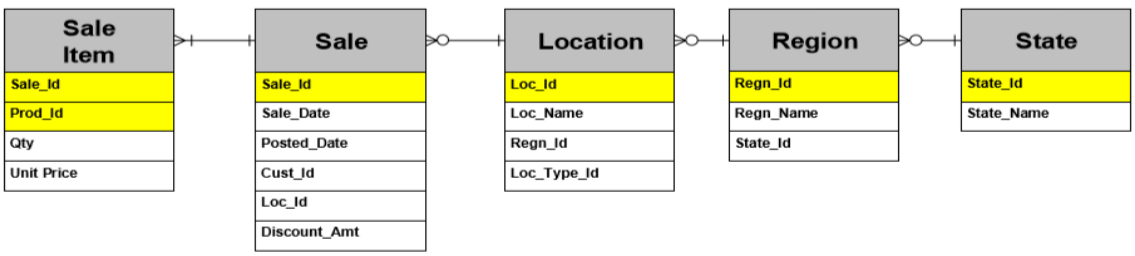
vulnerable for errors and beyond the capabilities of non-technical users. This is not a problem in transaction processing systems because the complexity of the database structure is hidden from the user by a layer of software (Moody and Kortink 2000, 3.)

Moody and Kortink (2000, 3) state that "a major reason for the complexity of operational databases is the use of normalization". Usually normalization multiplies the number of tables required. According to Codd (1970) the objective of normalization is to minimize excess data. However excessive data seems to be less of an issue in a data warehousing environment because data is not updated online (Moody and Kortink 2000, 3).

Given this premise of complexity and understanding, it is quite logical to try to separate the Meta data model design in central data warehouse environment and the data model used for data mart design. If the users are analysts and not equipped with deep knowledge of modelling it makes no sense for them to use time to learn it. It is better to focus on analyzing and making simple queries through the data mart. There are five main steps identified for this method.

1. Develop Enterprise Data Model (if one doesn't exist already)
2. Design Central Data Warehouse
3. Classify Entities
4. Identify Hierarchies (see figure 2-9)
5. Design Data Marts

Figure 2-9 – Example of hierarchy (Moody and Kortink 2000)



There are several benefits of this approach. First, it provides a more structured approach to developing dimensional models. Second, it ensures that the data marts and the central data warehouse reflect the underlying relationships in the data. Third, developing data warehouse and data mart designs based on a common enterprise data model simplifies extract and load processes. Fourth, an existing enterprise data model provides a useful basis for identifying information requirements in a bottom up manner, based on what data exists in the enterprise. Fifth, an enterprise data model provides a more stable basis for design than user query requirements, which are unpredictable and subject to frequent change. Sixth, it ensures

that the central data warehouse is flexible enough to support the widest possible range of analysis requirements, by storing data at the level of individual transactions. Aggregation of data at this level reduces the granularity of data in the data warehouse, which limits the types of analyses which are possible. Finally it maximizes the integrity of data stored in the central data warehouse.

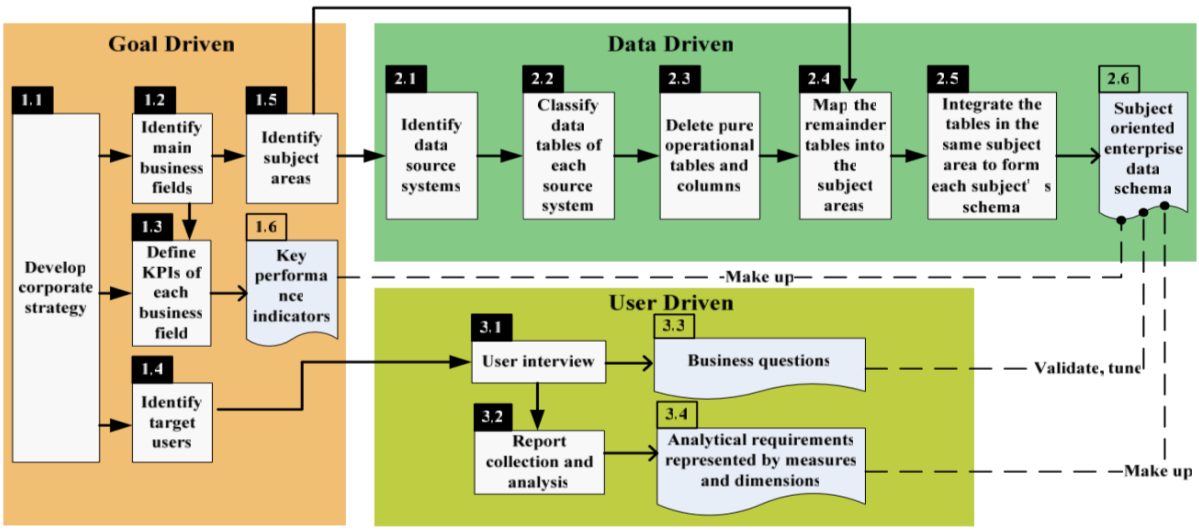
The approach provides much more guidance to designers of data warehouses and data marts than earlier approaches. However the approach will not be introduced in details in this thesis. Careful analysis is still required to identify the entities in the enterprise data model which are relevant for decision making. Thus, once this has been done, the development of a dimensional model should be rather easy. Using an entity relationship model of the data provides a much better starting point for developing dimensional models than starting from scratch, and can help avoid many of the pitfalls faced by inexperienced designers (Moody and Kortink 2000, 11.)

2.4.3 Triple-Driven Data Modelling Methodology by Guo et al. (2006)

Triple-Driven Data Modelling Methodology (title coined by Guo et al. 2006) describes a methodology for developing data warehouse logical data model (see appendix A2 for logical data model illustration). Guo et al. (2006, 59) criticize that data models using single fundamental approach are usually incomplete, and cannot obtain full satisfaction and trust of user and organizations simultaneously. Therefore their study tries to capture the essentials of data modelling process and combine the goal, data and user-driven approaches. There are four levels in the methodology: (1) goal-driven, (2) data-driven, (3) user-driven, and (4) combination level. The model takes into account all three approaches and combines them together.

The eventual result of combinations stage is a complete, subject-oriented logical data model of a data warehouse. The combination is based on the subject oriented data schema formed in the data-driven stage, making up the goal-driven and the user-driven results. In short, the goal-driven stage produces subjects and KPIs of important business fields. The data-driven stage obtains subject oriented data schema. The user-driven stage yields business questions and analytical requirements. The combination stage combines the triple-driven results.

Figure 2-10 – Triple-Driven Data Modelling Methodology (Guo et al. 2006)



Goal-Driven Level

1. Develop Corporate Strategy
2. Identify Main Business Fields
3. Define KPIs of Each Business Field
4. Identify Target Users
5. Identify Subject Areas
6. Deliverable: Key Performance Indicators

Data-Driven Level

1. Identify Data Source Systems
2. Classify Data Tables of Each Source System
3. Delete Pure Operational Tables and Columns
4. Map the Remainder Tables into the Subject Areas
5. Integrate the Tables in the Same Subject Area to Form Each Subject's Schema
6. Deliverable: Subject Oriented Enterprise Data Schema

User-Driven Level

1. User Interview
2. Reports Collection and Analysis
3. Business Questions
4. Deliverable: Analytical Requirements Represented by Measures and Dimensions

Combine the Triple-Driven Results

1. KPIs
2. Enterprise Data Schema
3. Analytical Requirements

In the figure 2-10 there are three parts in the logical data model. In fact, the three parts represent three different data layers of the same subject. The bottom layer is base data layer of the data-driven result part, which holds basic, crude and raw data collected from the operational systems. The medium layer is summary data layer of the user-driven result part,

which holds aggregate, statistical data around a subject. The high layer is advanced data layer of the goal-driven result part, which holds highly refined, deep computed data of a subject that executives, senior managers and decision-makers pay attention to. The three data layers are complementary to each other, providing a relatively complete data view of a subject (Guo et al. 2006, 65.)

There are several benefits of this approach. First, it ensures that the data warehouse reflects the enterprise's long term strategic goals and accordingly ensures actual business value of the data warehouse as well as stability of data model, which meets senior managers' expectations. Second, it raises acceptance and trust of users towards the data warehouse with users' involvements in the user-driven stage. Third, it leads to a design capturing all specifications. And finally it ensures that the data warehouse is flexible enough to support the widest range of analysis, by including the three different data layers: the data-driven base data layer, the user-driven summary data layer, and the goal-driven synthesis data layer.

The impacts of the methodology in research are encouraging. In one of the case studies, company started from a situation where operational databases were scattered and not integrated, and business needs and users' requirements were confusing. The proposed method was essential to direct them toward a solution that is both established in the data and oriented to business needs (Guo et al. 2006, 65.)

2.4.4 Comparison of Different Methodologies

Kimball's model does not combine the three different approaches. Instead of combining these approaches those are treated as separate methods. Guo et al. represents more advanced approach where all the three levels are combined. In future this kind of combination method is likely to provide the best probability of success because all the different ankles are taken into account. The classic data warehouse infrastructure follows the data model combo of Kimball, Moody and Kortink's. However the Guo's model seems to be quite comprehensive and it would be no surprise if it dominates Kimball's model in future. If Guo's and Moody and Kortink's methods were combined it would probably present the most optimal way to design data warehouse and data marts. In table 2-1 the main differences of alternative methodologies are listed. There is a linear developing slope from Kimball's model to the Guo's model if it is measured by depth.

Table 2-1 – Data modelling methodologies comparison

Kimball's Dimensional Modelling	Triple-Driven Data Modelling Methodology by Guo et al.	Data Warehouse and Data Mart Design by Moody and Kortink
Goal-driven	Goal-Driven	Develop Enterprise Data Model
Data-driven	Data-driven	Design Central Data Warehouse
User-driven	User-driven	Classify Entities
	Combine the triple-driven results	Identify Hierarchies
	Final results	Design Data Marts

2.5 Analytics in Data Warehouse Environment

Once the data model and architecture are finished and tested then the Data Warehouse is ready for the use of analytics. Objective of dimensional modelling is to produce database structures that enable end user to understand analyze and write query against it easily. The most used query language among analysts for this purpose is called SQL (Structured Query Language). A secondary objective is to maximize the efficiency of queries. It achieves these objectives primarily by minimizing the number of tables and relationships between them. This reduces the complexity of the database and minimizes the number of joins required in user queries (Moody and Kortink 2000, 3.)

There are two major differences between operational databases and data warehouses regarding the end user access and read-only view. In a data warehousing environment (OLAP), users write queries directly against the database structure, whereas in an operational environment (OLTP), users generally only access the database through an application system "front end". In a traditional application system, the structure of the database is invisible to the user. Data warehouses are effectively read only databases; users can retrieve and analyze data, but cannot update it. Data stored in the data warehouse is updated via batch extract processes (Moody and Kortink 2000, 2.)

From analytics point of view data warehousing represents the 1.0 era according to Chen, Chiang and Storey (2011, 1169). Core capabilities are ad-hoc queries, online analytical processing, reporting, dashboards, scorecards, interactive visualization, predictive modeling and data mining. In practice there are many solutions that can take the advantage of the structured data stored in data warehouses. Such software as IBM Cognos, Qlikview, Tableau and SAP are capable of simplifying the reporting task. These software are easy to use because you can create 'views' then include the items like sales report, charts and graphs that

are the most useful for specific purpose. One great feature is the time-dimensional filters and why not other filters too. These filters automatically update the whole field when the user clicks the preferred settings. If we compare these tools to Excel spreadsheets, the new tools are faster, easier and equipped with better visualization tools.

2.6 Data Warehouse Implementation, Development and Project Management

Implementing a data warehouse generally requires a massive effort from a company. Obviously large project always need to be carefully planned and executed according to established methods. The project lifecycle has many facets and no single person can do it all alone. Instead, implementing data warehouse requires expertise from various fields within a company. Data warehouse projects share the same basics with other information technology projects. Scope, time, cost, quality, human resources, communication, risk, procurement and integration are the items that require the management focus.

2.6.1 Success factors

Even though the fundamental project management is the same, there is a bunch of characteristics related to data warehouse projects. There are several lists developed by researchers, I am introducing a few of them in this section. Reeves (2009) and Solomon (2005) share a few guidelines regarding the critical questions that need to be asked during the planning process, some risks that should be weighed and some process to be followed to ensure a successful data warehouse implementation. According to Turban et al. (2011, 354) following their guidelines should increase the company's probability of success.

1. Establishment of service-level agreements and data-refresh requirements
2. Identification of data sources and their governance policies
3. Data quality planning
4. Data model design
5. ETL tool selection
6. Relational database software and platform selection
7. Data transport and conversion
8. Reconciliation process
9. Purge and archive planning
10. End-user support

According to Hwang and Xu's (2005) research data warehousing success is a multifaceted construct. Goal of improving user productivity should be kept in mind while building a data warehouse because it is likely to result in prompt information retrieval and

enhanced quality of information. Before Hwang and Xu, Weir (2002) wrote out his own recipe for developing data warehouse success. The recipe consists of ten items like the first guideline:

1. The project must fit with corporate strategy and business objectives
2. There must be complete buy-in to the project by executives, managers and users.
3. It is important to manage user expectations about the completed project
4. The data warehouse must be built incrementally
5. Adaptability must be built in
6. The project must be managed by both IT and business professionals
7. A business-supplier relationship must be developed
8. Only load data that have been cleansed and are of a quality understood by the organization
9. Do not overlook training requirements
10. Be politically aware

Wixom and Watson (2001) categorized these success factors to eight blocks: management support, champion, resources, user participation, team skills, source systems and development technology. There are several of success factor lists like these published, but the three lists mentioned above include the major elements represented in the research.

2.6.2 Risk factors

According to Turban et al. (2011, 354) data warehouse risks are more serious because data warehouses are large and expensive projects. Adelman and Moss (2001) identify a large set of risks.

1. No mission or objective
2. Quality of source data unknown
3. Skills not in place
4. Inadequate budget
5. Lack of supporting software
6. Source data not understood
7. Weak sponsor
8. Users not computer literate
9. Political problems or turf wars
10. Unrealistic user expectations
11. Architectural and design risks
12. Scope creep and changing requirements
13. Vendors out of control
14. Multiple platforms
15. Key people leaving the project
16. Loss of the sponsor
17. Too much new technology
18. Having to fix an operational system
19. Geographically distributed environment
20. Team geography and language culture

Naturally the sooner the risks are identified the sooner the company can try and mitigate them. Turban et al. (2006) agrees on some of these points. They point out the following factors that are a great threat for the success: cultural issues being ignored, inappropriate architecture, unclear business objectives, missing information, unrealistic expectations, low levels of data summarization and low data quality.

2.6.3 Points of failure

Turban et al. (2011, 355) identify several issues when developing data warehouse. If the company fails to plan the data warehouse project carefully, one can say it is planning to fail. For example if the consultants or whoever is leading the execution is not managing expectations, the project is doomed to fail. When we measure success, it usually is achieved when expectations are met. Such a simple task can become a showstopper. This list includes the mistakes that company should avoid.

1. Starting with the wrong sponsorship chain
2. Setting expectations that you cannot meet
3. Engaging in politically naïve behavior
4. Loading the warehouse with information just because it is available
5. Believing that data warehousing database design is the same as transactional database design
6. Choosing a data warehouse manager that is technology oriented rather than user oriented
7. Focusing on traditional internal record-oriented data and ignoring the value of external data and of text, images and perhaps sound and video
8. Delivering data with overlapping and focusing definitions
9. Believing promises of performance, capacity and scalability
10. Focusing on ad hoc data mining and periodic reporting instead of alerts

Many of the points are easy to understand but the wider perspective control is hard to obtain. But for instance why should data warehouse manager be user oriented? It is so that the data warehouse is built to help the analysts and other consumer of the data. Of course companies need the technical skills, but according to Turban et al. (2011, 355) it is more important to let users get what they need, “not just advanced technology for technology’s sake”. The last point might seem a little confusing. It means that users like managers may want to prioritize their tasks and not do the data mining, which belongs to analysts. Managers can set up alerts in the monitoring process and the data warehouse would inform them if there are abnormal variations or events in the data. This kind of method naturally saves some time for managers.

2.6.4 Incremental and Radical Change in Technology Adoption

The change in organization and business processes can be divided into two different categories, incremental or radical. As mentioned earlier in success factor lists, data warehouses should be implemented incrementally. Incremental change is continuous path starting from existing processes, proceeding step-by-step to its goal by carrying only moderate operational risk. In turn, radical change is done once from clean table by carrying high operational risk. These two approaches are used in very different situations. For example, incremental change is better when company needs to cut costs or change their existing process to be more efficient, improve processes. Different elements of the incremental and radical change are shown in figure 2-11 (Saarinen, 2013).

Figure 2-11 – Comparison among incremental and radical change (Saarinen 2013)

	Improvement	Innovation
Level of change	Incremental	Radical
Starting Point	Existing Process	Clean Slate
Frequency of Change	One-time/continuous	One-time
Time Required	Short	Long
Participation	Bottom-up	Top-Down
Typical Scope	Narrow, within functions	Broad, cross-functional
Risk	Moderate	High
Primary Enabler	Statistical Control	Information Technology
Type of Change	Cultural	Cultural/Structural

Incremental Change

The incremental change model is linked to the sociotechnical change approach, where the change is influenced both hard and soft system changes. Additionally, Cooper et al. (1995) describes that incremental change works better in humane change projects, where company need to train personnel and give personnel keys to do the needed change in processes by themselves. In incremental change model, employees who are recipients of the change must work and implement the change. Employees and leadership of existing process are used in change management and the communication about the change is wide and open for all. The speed of the change is determined by the capabilities of existing employees, thus milestones are flexible. Therefore, the pace of this type of change should be comfortable for

existing employees and to all other internal and external constraints which company has at the point of change.

The motivation for the change comes from internal dissatisfaction for existing processes. Before IT takes place and consolidate new processes, these new processes are stressed by piloting them. Additionally, incremental change model assumes that change is most suitable in tiny steps at a time (Jarvenpaa and Stoddard, 1998.) Ettlie et al. (1984) states that incremental change adoption tends to happen more likely within large, complex and decentralized companies that dominates markets with their growth strategies.

The major advantage of the incremental model is that the general risk of failure is small because many existing employees participate in the change, thus each employee can feel ownership for the changes happening. In general, incremental model increases company's capacity for change. Additionally, translating radical vision into multiple incremental targets helps the company to get started with the change project which could otherwise be seen unreachable (Jarvenpaa and Stoddard, 1998.)

The major disadvantage of the incremental model is the long time span to accomplish the vision, the vision that should be alive and reminded to employees every once and while, even though the market conditions change. Otherwise, the company can lose their sight into the motivation for their radical vision. The danger is that company declares victory too soon, after modest changes and turns their sight into newer focus points (Jarvenpaa and Stoddard, 1998.)

Radical Change

In receipt of successful change, the radical change model is often linked to gradual steps which change the deep structure of the company. Radical change can also be sudden, revealed quickly and amend essentially the basic assumptions, business processes, culture and the structure of company. The change is easier if company faces identity crisis and disorder. The participation of the change must be top-down and lead by the CEO (Jarvenpaa and Stoddard, 1998.) In addition, senior management must motivate employees by sharing a common vision and creating appropriate culture as well as developing requisite internal alliances (Nadler, et al, 1995; Ettlie et al. 1984). External resources and outside vision is required to succeed. Persons outside the company, without fear of challenging existing processes are hired to lead and participate in the change. These persons can be consultants or executives new to the company or process that is being re-engineered. They might be also

from other parts of the company, who have no earlier knowledge of the processes under the change (Jarvenpaa and Stoddard, 1998.)

Ettlie et al. (1984) states that centralization of decision making increases in radical change projects. Therefore, the change team should be tiny but devoted. The communication about the becoming change should be limited and only in a level of have-to-know basis. Motivation behind the change arises from internal crisis and milestones are sharp to be concise when the old is replaced with new ways of doing things. The radical change process target usually for new advanced IT and therefore qualifies all employees for the new process (Jarvenpaa and Stoddard, 1998.)

Jarvenpaa and Stoddard (1998) mentioned four conditions for companies that succeed with fast evolving radical change. First, company needs to have a real performance crisis. Second, the change must take place in a tiny self-contained unit. Third, company or parent need to have lots of money to cover fast evolving radical change. Finally, companies need to have ability to borrow and replant solutions like buying software packages from outside.

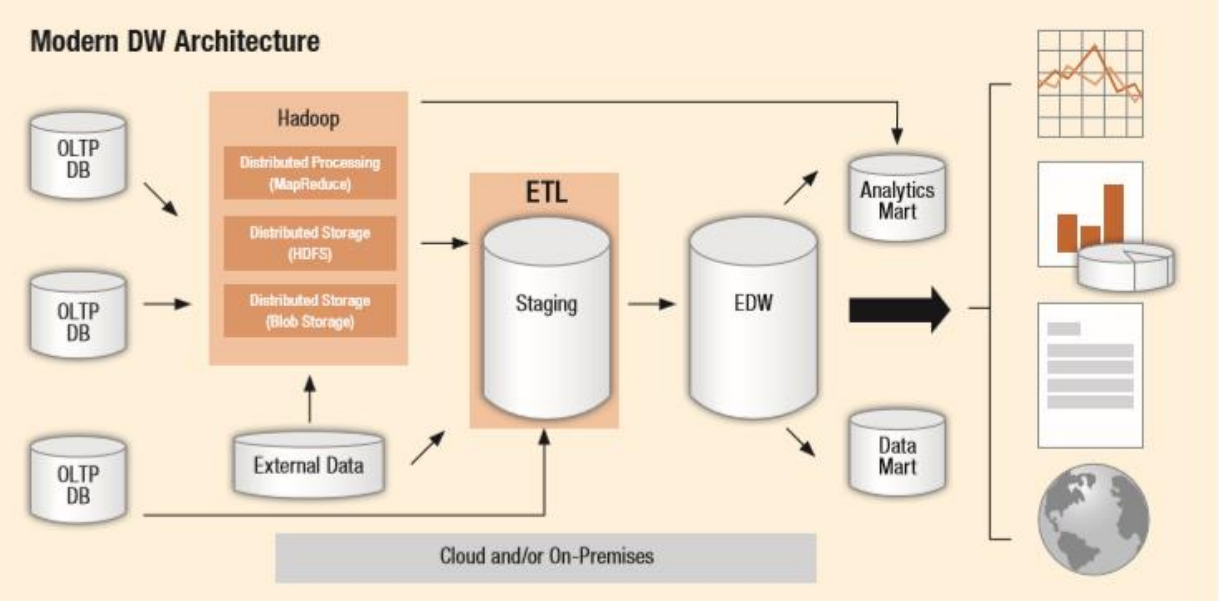
Most of the good managers hate the radical model because radical model challenge much of managing and motivating employees. Additionally, the radical change insists managers to cannibalize their own business. Some employees are naturally left out and their insecure position in company form bottlenecks to prevent the change. To create need for change, the change team uses reverse values that most companies have institutionalized: empowerment, self-management, and innovation from bottom-up. Therefore, the change team requires robust control and daily personal involvement from top management. For that reason, top management have only limited time to spend into following fast evolvement of marketplace, potentially leading to lost market opportunities or misaligned strategy (Jarvenpaa and Stoddard, 1998.)

The major advantage of the radical model is that the change is accomplished quickly (Jarvenpaa and Stoddard, 1998). Radical change stands for heroism and tough decision-making, for example, cost cuttings, downsizing and changing structure of company (Nadler et al 1995). On the other hand one of the major disadvantages of the radical model is that it increases risks in change project (Nadler et al 1995). If radical change fails, it can lead to chaos, and company and individuals may lose their identity (Gersick 1991; Clemons 1995).

2.7 Modern Data Warehousing Architecture

According to D'Antoni and Lopez (2014, 9) modern BI&A and big data wave has moved beyond the "only for Web start-ups" or "only for scientific use" phase and is now ready to answer real-world business questions. The modern enterprise data warehouse (EDW) architecture (see figure 2-12) needs to bring together the technologies and data required to support traditional business needs and stronger predictive analytics, leveraging large data sets.

Figure 2-12 – Modern Data Warehouse Architecture (D´Antoni and Lopez 2014)



Dull (2014, 7) gives companies three options to cope with growing volumes of data:

1. Add more hardware and/or horsepower to their existing EDW and operational systems (viable but expensive).
2. Consider alternative ways to manage their data (Use Hadoop as staging platform).
3. Do nothing (kiss of death for some)

In order to tame the big data, companies need to modernize their classic data warehouse architecture. Compared to classic data warehousing architecture, the modern architecture contains such building blocks as Hadoop technology, Apache Pig tools for ETL process and Analytics Mart for advanced analytics.

2.7.1 Hadoop Technology

Hadoop is the technology developed, sponsored and supported by Hortonworks with the most promising potential in the big data space and it started simply as a project at Yahoo! to build a better search engine and process all that data. Since then it has evolved into the

centerpiece of modern data analytics architecture, with a large group of open source components surrounding it. Hadoop's power quickly brought it to the fore for large-scale data processing. Hadoop got two core elements: a framework for data processing called MapReduce and a distributed file system known as the Hadoop Distributed File System (HDFS). These technologies combine to allow massive parallelism and fault tolerance while running on commodity hardware.

1. MapReduce is "the resource management and processing component of Hadoop. MapReduce allows Hadoop developers to write optimized programs that can process large volumes of data, structured and unstructured, in parallel across clusters of machines in a reliable and fault-tolerant manner. For instance, a programmer can use MapReduce to find friends or calculate the average number of contacts in a social network application, or process web access log stats to analyze web traffic volume and patterns" (Dull 2014, 6).

2. The Hadoop Distributed File System is "the data storage component of the open source Apache Hadoop project. It can store any type of data: structured, semi-structured and unstructured. It is designed to run on low-cost commodity hardware and is able to scale out quickly and cheaply across thousands of machines" (Dull 2014, 5). There are two benefits with this approach: *Storage costs:* due to low cost of Hadoop storage, you could store both versions of the data in the HDFS: the before application data and the after transformed data. All of the data would then be in one place, making it easier to manage, reprocess and analyze at a later date. *Processing power;* processing data in Hadoop frees up EDW resources and gets data processed and transformed into your EDW quicker so that the analysis work can begin.

D'Antoni & Lopez (2014, 10) agree that the cost of Hadoop storage is lower than e.g. SAN (storage area network) storage: A common motto in modern computing is that storage is cheap but this is far from the case with large enterprises that utilize storage area network (SAN) for storing. The average cost for enterprise SAN storage was $4,876 per terabyte in 2011 (Guevara et al., 2011). Even when allowing for some reduction in cost over time, storage is a major part of IT's ongoing operating expense. We can use an analytic architecture that is optimized to process larger data volumes to leverage costs and benefits of storage and processor budgets appropriately. (D'Antoni & Lopez 2014, 10)

To cope with the costs of storing Hadoop got three advantages. First, scale out instead of up. In the relational data warehouse environment, performance is usually improved by

using larger and faster hardware (which tends to be also exponentially more expensive as it grows in scale). In the Hadoop world, more nodes also known as servers are added and the work is done in parallel. Second is commodity hardware. Hadoop is designed around dense, local storage and large sequential reads. Third is parallel processing. Hadoop is designed to manage and support massively parallel processing (MPP), which is optimized for processing very large data sets (D'Antoni & Lopez 2014, 10.) Hadoop is a free open source project, but most organizations will choose to go with a commercial distribution for ease of management. The annual licensing cost for the commercial solutions is about $4,000/node/year; however that is not insignificant but is significantly lower than the cost of a commercial RDBMS that can be as high as $50,000 per core (Bantleman, 2012).

One advantage of Hadoop is that data can be stored in its raw, native state. It does not need to be formatted upfront as with traditional, structured data stores; it can be formatted instantly upon the data request. This process of formatting the data at the time of the query is called "late binding" and is a growing practice for companies. Late binding ensures that there is context to the data formats depending on the data request itself. Thus, Hadoop programmers are able to save months of programming by loading data in its native state (Dull 2014, 16.)

On a side note you don't need big data to take advantage of the power of Hadoop even though it seems to be popular belief. Not only can you use Hadoop to ease the ETL burden of processing your "small" data for your EDW you can also use it to offload some of the processing work you're currently asking your EDW to do. You can simply use Hadoop to update data in your EDW and/or operational systems. In short: "send the data to be updated to Hadoop, let MapReduce do its thing, and then send the updated data back to your EDW. This would not only apply to your EDW data, but also any data that is being maintained in your operational and analytical systems. Take advantage of Hadoop's low-cost, high-speed processing power so that the EDW and operational systems are freed up to do what they do best" (Dull 2014, 9).

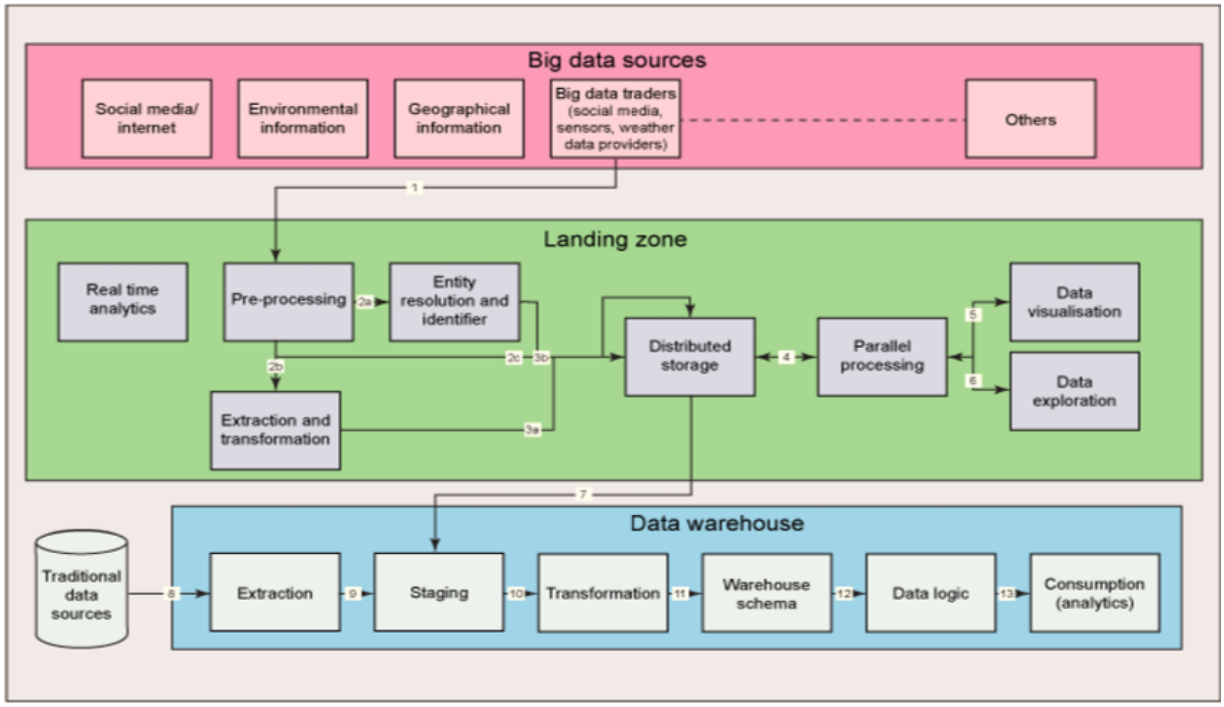
2.7.2 Analytics Mart

A common use case is to build a model for predictive analytics and run it against real-time data. These models will be built over and over again and run many times in an effort to perfect the models, so service times for these solutions must be very good. Hadoop HDFS has not been the platform for these real-time analytics; a more common scenario is to extract data

from HDFS and load it into a memory-optimized columnar platform that allows for a high degree of data compression. Many columnar databases still support SQL and offer scale-out MPP (massively parallel processors) on a similar hardware platform to HDFS (D'Antoni & Lopez 2014, 10.)

Traditional data warehousing is focused on operational metrics such as stock, supply chain, and operational goals. These metrics tend to look at historical and current data, and although they may allow for some forward-looking forecasting, they usually look at internal data only, with limited use of outside data sources (D'Antoni & Lopez 2014, 10.) With years of evolution and ever more powerful hardware, data warehouses have become repositories allowing for large-scale reporting and analysis. According to Jain and Nandi (2014, 5) Hadoop is supporting the data warehouse as source of data (see figure 2-13).

Figure 2-13 – Big data as a data source to the modern data warehouse (Jain & Nandi 2014)



2.7.3 Data Modelling in Modern Data Warehousing

In addition to the modelling efforts described in the classic data warehouse architecture, data architects can provide value to the Hadoop tasks as well. According to D'Antoni and Lopez (2014, 13) Data models for OLTP/OLAP systems will still be required when that data is used in Hadoop and data models should be prepared for external data sources from where the data is brought in. Data architects can assist in the design of for example Hive Query Language (HiveQL) "tables".

Apache Hive refers to itself as "data warehouse software which facilitates querying and manages large data sets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL". According to D'Antoni and Lopez (2014, 10) this language is based on SQL, therefore developers and analysts can more easily query HDFS data via language they are used to. When a user runs a query in HiveQL, a MapReduce job is generated and launched to return the data and therefore no Java coding is required. Data models of the physical file store in Hadoop (HDFS) aren't required, but logical data models of the data that is managed there for any length of time would be. Many modern data modelling tools have begun to support Hive schemas for these reasons. These tables can then be documented along with all the other enterprise data assets.

2.7.4 Data Warehouse Modernization

IBM (2015) states that data warehouse modernization also known as data warehouse augmentation is about building on an existing data warehouse infrastructure and leveraging big data's potential technologies to augment its capabilities. There are three key types of data warehouse modernizations according to IBM (2015, direct quote):

1. Pre-Processing - using big data capabilities as a "landing zone" before determining what data should be moved to the data warehouse

2. Offloading - moving infrequently accessed data from data warehouses into enterprise-grade Hadoop

3. Exploration - using big data capabilities to explore and discover new high value data from massive amounts of raw data and free up the data warehouse for more structured, deep analytics.

The Data Warehousing Institute (TDWI) provides assessments and models for companies to measure their maturity of adapting advanced analytics and big data. Advanced analytics maturity assessment is introduced in the case study. IBM is also sharing their short questionnaire to find out if the company needs data warehouse modernization: "1. Are you integrating big data and data warehouse capabilities to increase operational efficiency? 2. Have you taken steps to migrate rarely used data to new technologies like Hadoop to optimize storage, maintenance and licensing costs? 3. Are you using stream computing to filter and reduce storage costs? 4. Are you leveraging structured, unstructured, and streaming data sources required for deep analysis? 5. Do you have a lot of cold or low-touch data that is

driving up costs or slowing performance? If you answered yes to any of the above questions, the data warehouse modernization use case is the best starting point for your big data journey."

With data warehouse modernization, organizations can combine streaming and other unstructured data sources to existing data warehouse, optimize data warehouse storage and provide query accessible archive, rationalize the data warehouse for greater simplicity and lower cost, provide better query performance and quality to enable complex analytical applications and deliver improved business insights to operations for real-time decision making (IBM 2015.) Challenges in big data implementation are related to changing hardware and software in addition to data backups. These have never been easy or inexpensive for IT organizations as evidenced by the large number of firms still using mainstream platforms. When considering large amounts of data, backups are always a challenge. Given the nature of HDFS and the challenges of backing up massive data volumes, many firms choose to forego performing backups of these data volumes, which could leave them vulnerable in a disaster. There are options from some Hadoop vendors for disaster recovery if your organization needs it for its analytic platform (D'Antoni and Lopez 2014, 13.)

From human resource perspective, your organization needs the following key abilities: Linux system administrators, automation engineers, Java and data analysis. Compared to a traditional model, where the database administrator (DBA) manages the data warehouse database, the DBA role does not apply in HDFS. Linux system administration skills are very important, because the majority of implementations are running on Linux platforms, where community support is widely available. When dealing with tens or thousands of cluster nodes, automation becomes very important. Software and firmware updates are also candidates for automated processing (D'Antoni and Lopez 2014, 14.)

Leveraging cloud computing for big data makes for an interesting solution particularly if workloads vary a lot. Like other cloud computing offerings, there are usually two types of solutions: platform as a service (PaaS) a.k.a. Hadoop as a service (HaaS) and infrastructure as a service (IaaS). Most major cloud vendors have HaaS offerings which can be a simple way to get up and running with Hadoop and the toolkit overnight. This means that the vendor manages the entire underlying infrastructure and you manage the configuration and of Hadoop. The IaaS offerings like IBM Cognos involve spinning up a number of virtual machines (VMs) and building a Hadoop cluster on them. This places more

configuration work onto your staff but provides more flexibility with the tools installed alongside Hadoop (D'Antoni and Lopez 2014, 14.)

Getting large, existing data volumes into the cloud is a major issue. The good news here is that most cloud providers do not charge a fee to upload data. Like most other cloud computing solutions, the benefits involve flexibility and low initial capital investment. Even though firms are concerned about security when moving to a cloud computing model cloud providers are going out of their way to address these concerns (D'Antoni and Lopez 2014, 14.)

Big data platforms are not totally free, but as mentioned earlier there are some clear cost advantages which are good for the company's budget. According to D'Antoni and Lopez (2014, 15) "Because performance is achieved through horizontal scaling and aggregate resources, individual nodes do not need to be as powerful as a monolithic server". Hadoop and most other big data and NoSQL (not only SQL) platforms leverage dense, local storage that comes at a much lower cost than enterprise SAN storage. All of these software platforms run nearly exclusively on Linux and most implementations take place on completely free distributions of the operating system (D'Antoni and Lopez 2014, 15).

Finally the data warehouse professionals need to understand that Hadoop and other big data technologies are not an either-or decision. Every design decision comes down to cost, benefit, and risk. During these times companies have the opportunity to leverage these new sets of technologies with only slightly modified data warehouse architecture. In order to leverage a plethora of data sources companies need to open their minds to the modern data warehouse architecture (D'Antoni and Lopez 2014, 15.)

2.8 Advanced Analytics in Modern Data Warehouse Architecture Environment

In past data management and warehousing were considered the foundation of BI&A 1.0. Design of data marts and tools for extraction, transformation, and load (ETL) are essential for converting and integrating enterprise-specific data. Database query, online analytical processing (OLAP), and reporting tools based on intuitive, but simple, graphics are used to explore important data characteristics Currently the BI&A 2.0 & 3.0 are focused on web-based unstructured, mobile and sensor-based content (Chen et al. 2012.)

Although MapReduce is a powerful and robust framework, writing Java code in mass scale would have required retraining data analysts and other IT personnel, who are used to working with SQL and scripting. In this way SQL is easier to learn and use. This skills and tools mismatch meant that enterprises were unlikely to adopt Hadoop solutions. The open source community realized these limits and brought together several projects such as Hive, Pig, and later Impala to provide a more user-familiar interface to HDFS (D'Antoni and Lopez 2014, 10.)

As mentioned earlier Hive functions as a SQL main data store on top of HDFS. This language is based on SQL, so developers and analysts can more easily query HDFS data. Apache Pig also builds a high-level procedural language that acts as an interface to HDFS. Pig is more frequently utilized in ETL scenarios than for just returning data results. Pig uses a text-based language called Pig Latin, which focuses on ease of use and extensibility. Apache Impala is part of a number of second-generation Hadoop solutions (along with Spark and Shark) that leverage memory-based processing to perform analytics. Impala has access to the same data in the HDFS cluster (and typically relies on the Hive metastore for table structures) but it doesn't translate the SQL queries that it is processing into MapReduce. Instead, Impala uses a specialized distributed query engine similar to those found in commercial parallel relational database management systems (D'Antoni and Lopez 2014, 10.)

2.8.1 Advanced Analytical Tools' Vendors Overview

Apart from the Hadoop technology and the peripheral solutions the market for other tools is under fierce competition. Gartner's view is that the market for BI and analytics platforms will remain one of the fastest-growing software markets. The market grew 9% in 2013, and is projected to grow at a compound annual growth rate of 8.7% through 2018. Gartner has evaluated most of these tools in their latest research during February 2015 when they surveyed 2,083 users of BI platforms. Vendors are assessed for their support of four main use cases: 1. Centralized BI: Supports a workflow from data to IT-delivered and managed content. 2. Decentralized Analytics: Supports a workflow from data to self-service analytics. 3. Governed Data Discovery: Supports a workflow from data to self-service analytics to systems-of-record, IT-managed content with governance, reusability and promotability. 4. OEM/Embedded BI: Supports a workflow from data to embedded BI content in a process or application. Original Equipment Manufacturer (OEM) means that some companies may use for example Qlik-products as part of their own product to deliver value to their customers. Vendors are also assessed according to the following 13 critical

capabilities: business user data mashup and modelling, internal platform integration, BI platform administration, metadata management, cloud deployment, development and integration, free-form interactive exploration, analytic dashboard and content, IT-developed reporting and dashboards, traditional styles of analysis, collaboration and social integration and embedded BI (Sallam et al. 2015.)

Figure 2-14 – Evaluation of analytical tools (Sallam et al. 2015)

Leaders (Tableau, Qlik, SAP, SAS, IBM, Microsoft, MicroStrategy, Oracle, Information Builders)

Leaders can deliver on enterprise wide implementations that support a broad BI strategy. Leading vendors are strong in the width and depth of their BI platform capabilities, delivering an excellent customer experience, product vision, innovation, market growth and momentum, marketing and sales differentiation and effectiveness, and having capabilities that are used broadly (Sallam et al. 2015.)

Challengers (Birst, Logi Analytics)

Challengers are well-positioned to succeed in the market. However, they may be limit themselves to specific use case or technical environments. Their complex vision slow them down due to lack of coordinated strategy across the various products in their platform portfolios, or they may lack the marketing efforts, geographic presence, industry-specific content and awareness compared to the vendors in the leaders quadrant (Sallam et al. 2015.)

Visionaries (Tibco, Alteryx, Panorama Software)

Visionaries obviously have a strong and unique vision for delivering a BI platform. They offer special and in-depth functionality in the areas they address. However, they may have gaps relating to broader functionality requirements. Visionaries are thought-leaders and innovators, but they may be lacking in scale, or there may be concerns about their ability to grow and provide consistent execution (Sallam et al. 2015.)

Niche Players (Prognoz, Pentaho, GoodData, Yellowfin, Datawatch, Pyramid Analytics, Targit, Board International, Salien Management Company, OpexText)

Niche players do well in a specific segment of the BI&A platform market, such as collaboration, reporting, dashboards, or big data integration. They are likely to lack depth of functionality elsewhere. They may also have gaps relating to broader platform functionality or have less than stellar customer feedback. Niche players may have a reasonably broad BI platform, but limited implementation, resource and support capabilities or relatively limited customer bases. In addition, they may not yet have achieved the critical mass or necessary scale to solidify their market positions (Sallam et al. 2015.)

Tableau and Qlik are excelling at delivering on current market and customer experience requirements. They are satisfying customers for data discovery in addition to easy and broader use. The vendors with the majority of the market momentum are focused on making it easier and simpler for more users to author content and explore and discover patterns in data wherever they are. Tableau and Qlik are growing from new analytics project investments. Qlik has recently introduced its new Qlik Sense platform, while Tableau is adding elements incrementally on each new platform release. One significant difference between Qlik and Tableau is the cost. Tableau's pricing seems to be way higher than Qlik's (Sallam et al. 2015.)

SAP, SAS and IBM are investing aggressively to regain momentum and differentiation through a smart data discovery experience. They are also positioning their integration with their enterprise platforms to support governed data discovery as key differentiators. SAS has had better traction, adoption and customer experience than IBM and SAP as a result of its major commitment to SAS visual analytics, its data discovery capabilities. SAP also has a more logical road map than IBM. IBM has a compelling vision for Watson analytics combining self-service data preparation, natural-language query generation and exploration, automatic pattern detection and prediction that will likely drive future market requirements. However, its road map for how this capability will integrate with and breathe momentum back into IBM Cognos is less clear (Sallam et al. 2015.)

Microsoft, Oracle, MicroStrategy, and Information Builders are slowed down by execution issues. Microsoft has delivered data discovery capabilities in Excel spreadsheets that have had some level of adoption. It has a strong product vision, particularly with self-service data preparation. However, mediocre product scores and the lack of a strong BI and analytics marketing and sales focus has limited Microsoft's market penetration and position to date (Sallam et al. 2015.)

From top vendors' perspective it seems that Tableau, Qlik, SAS, SAP and IBM are leading the way in advanced analytics' tools market. Depending on the requirements, company should choose either Tableau or Qlik if they are looking for easy configuration and easy to use platform which allows self-exploration. On the other hand SAS, SAP and IBM are at the moment the strongest with smart data discovery and automated pattern detection.

2.8.2 Analytics' Maturity

Before rushing into the advanced analytics, companies should evaluate their aptitude and maturity for taking advantage of the new methods. If a company is not ready, the time and money are wasted because the desired benefits might be impossible to reach if there is for example outdated data infrastructure. Halper and Stodder (2014, 11) suggest that analytics maturity can be described as the evolution of an organization to integrate, manage, and leverage all relevant internal and external data sources into key decision points. This basically means that companies can create an ecosystem that enables business insights and actions. Analytics maturity is not only about having some technology in place; it involves technologies, data management, analytics, governance, and organizational components. It can take years to create an analytics culture in an organization (Halper and Stodder, 2014).

A maturity model for analytics seems to be useful for any enterprise considering or in the process of implementing an analytics project. First, it helps create structure around an analytics program and determine where to start. Secondly, it also helps identify and define the organization's program goals and creates a process to communicate that vision across the entire organization. A maturity model will provide a methodology to measure and monitor the state of the program and the effort needed to complete the current stage, as well as steps to move to the next stage of maturity. It serves as a kind of odometer to measure and manage your progress and adoption within the company for an analytics program (Halper and Stodder, 2014.)

2.8.3 Trends in Analytics

As mentioned in the introduction part many trends in analytics are relevant for companies looking to become more mature in their analytics efforts. Halper and Stodder (2014) identify eleven trends that will be explained in detail below. These include:

1. Ease of use. Vendors that are providing analytics tools have made user interfaces easier to use, or even drag and drop. SQL-based software is quite straightforward for collecting data. Preparing data and visualizations have become easier to construct. Some vendors provide new ways to bring data together, such as data blending, where the data is combined without integrating it into a data warehouse whereas some vendors provide automation techniques for more advanced analytics where the software actually suggests a model using the outcome variables and an examination of the data. Ease of use is important in analytics maturity because it allows democracy and scalability in the use of data. Thus it increases the odds to become more successful and data driven (Halper and Stodder 2014, 6.)

2. The democratization and consumerism of analytics. As mentioned above ease of use is the move to make analytics available to more people in the organization. From the executive and business leader level to production units, users increasingly depend on data and analytics for all kinds of decisions. Many organizations would like to enable democracy in BI and analytics. This would allow a broad range of non-IT users to do more on their own with data access and analysis in other words it would be like a self-service model. Self-service BI and visual data discovery technologies are increasingly becoming more popular in enabling users to develop more sophisticated analytics and execute queries themselves. IT's role would be focusing on securing the governance and guidance (Halper and Stodder 2014, 6.)

3. Business analysts using more advanced techniques. Ease of use is also linked to the move from the statistician/modeler to a new user of predictive analytics. With the help of training business analysts are slowly becoming the new users of advanced analytics techniques such as predictive models. Business analysts might build relatively simple and straightforward models and this would free the data scientist (typically a scarce resource) to build more complex and sophisticated models (Halper and Stodder 2014, 6.)

4. Newer kinds of analytics. In addition to predictive models, other kinds of analytics are emerging to help drive business value. These include text analytics (unstructured text), social media analytics, geospatial analytics (GPS data), and clickstream analysis (behavior on websites). All of these techniques are starting to become more mainframe and can provide important insight, either by themselves or in combination with other techniques. The more mature an organization is in its analytics efforts, the more it makes use of newer forms of analysis (Halper and Stodder 2014, 7.)

5. Operationalizing analytics. Once something is operationalized, then it is part of a business process. Making analytics real is important because it helps make analytics more actionable and hence drive more value. For example a data scientist might build a predictive model, once the model is embedded in a system, only then it can create scalable value to the company. Operationalizing analytics helps make it more consumable, which is one of the core objectives of analytics development (Halper and Stodder 2014, 7.)

6. Big data. An important point about big data is that it is helping to drive the use of already existing techniques (like DW) as well as the development of new techniques for data analysis. Big data is driving the use of newer infrastructure such as Hadoop and DW environments that manage, process, and analyze new forms of big data, non-structured data, and real-time data. This might include NoSQL databases, DW appliances, and relational databases (Halper and Stodder 2014, 7.)

7. New development methods. Many organizations are employing agile methods. These faster and incremental cycles have helped organizations toward better collaboration between business and IT, faster and more iterative development cycles, and ultimately higher quality and satisfaction (Halper and Stodder 2014, 7.)

8. Open source. Open source is rapidly becoming state of art solution for infrastructure as well as analytics. Hadoop is a great example of how open source technologies are unlocking value in analytics. There is a whole ecosystem of tools and

techniques that have been developed to make the Hadoop Distributed File System (HDFS) more user-friendly. Open source is important because it enables crowdsourcing and community to innovate (Halper and Stodder 2014, 8.)

9. The cloud. Although it has taken longer than some expected for the cloud to be used in BI, it is now starting to become mainstream. One reason organizations are trying to move toward cloud is to offset costs with zero capital expenditure on infrastructure, maintenance, and even personnel. Cloud could solve the challenges of BI cost-efficiency and long time to deploy (Halper and Stodder 2014, 8.)

10. Mobile BI and analytics. The increasing adoption of mobile devices has opened up new ways to access and consume data and analytics anywhere anytime. To address security, performance, and availability concerns, some organizations may deploy cloud services to provide BI and analytics platform support for mobile users (Halper and Stodder 2014, 8.)

11. Analytics platforms. More companies are adopting analytics platforms with integrated solution for analytics. This includes data management, data preparation, and data analysis capabilities. The platforms can drive efficiencies into the analytics life cycle because they can help bring together the data as well as analyze it. The platform can be delivered in different ways: in the cloud, on premises, as an appliance, or in an integrated solution (Halper and Stodder 2014, 8.)

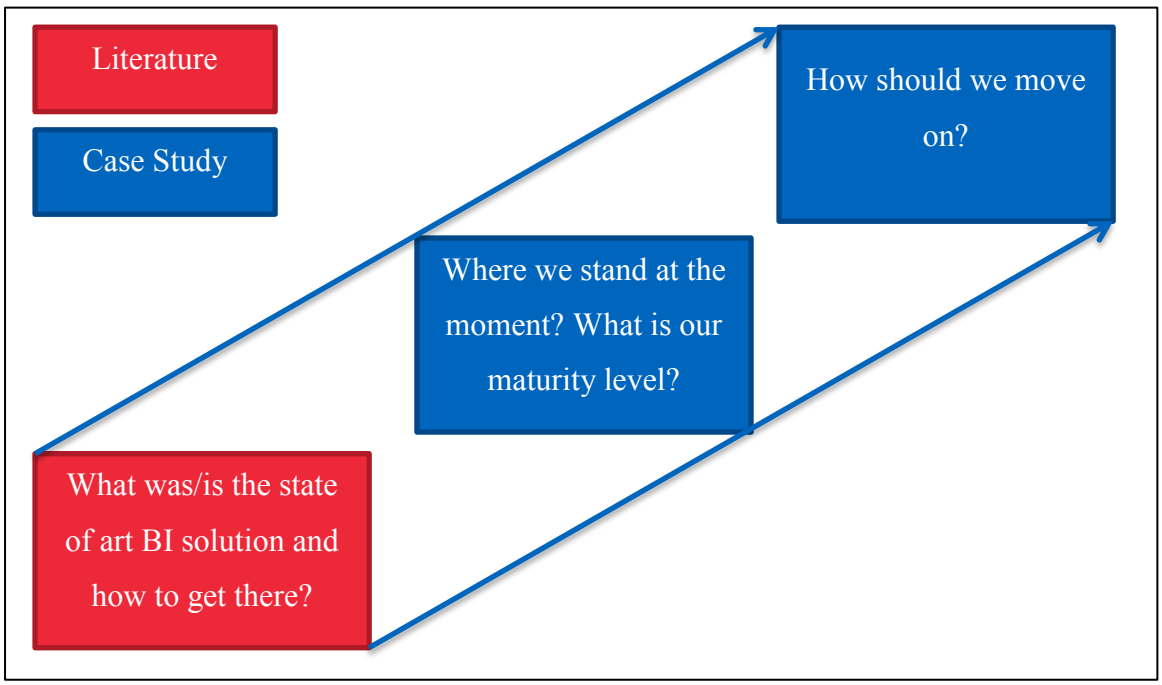
3 METHODOLOGY

This chapter describes research approach, data collection methods, respondents and finally the analysis part. In other words the point of methodology part is to illustrate the drawing of two islands, theory and practice, and a bridge between them. The theory introduces the current best practices to move on with analytics and the practice would be the status quo in the case company that is benchmarked against the theory with the help of survey and by conducting analytical overview of current data infrastructure.

3.1 Research Approach

The purpose of this study is to deliver a future analytics strategy for the case company that aims to build strong analytics culture and wants to be truly data-driven when it comes to all decisions and decision making processes. The case company will remain anonymous throughout the whole study due to sensitive nature of gathered information. In order to develop a future strategy both theory and practice play an important role.

Figure 3-1 – Research approach illustrated



This study is an exploratory research which starts with a literature review and is followed by a case study which consists of survey and BI architecture analysis. As usually in exploratory studies the data in this study is also qualitative. Qualitative research uses text as empirical material instead of numbers, is interested in perspectives of participants, in everyday practices and everyday knowledge referring to the issue under the study (Flick

2007). According to Sue and Ritter (2012, 2) the purpose of exploratory research is to formulate problems, clarify concepts, and form hypotheses. In this research the purpose is to suggest a new strategy once the concept and problem are clarified. In order to conduct a valid exploratory case study with good quality there are three factors that needs to be taken into account (explanation / phase):

1. Construct: identifying correct operational measures for the concept being studied/data collection
2. External: defining the domain to which a study's findings can be generalized/research design
3. Reliability: demonstrating that the operations of a study such as the data collection procedures can be repeated, with the same results / data collection

In this thesis I used a case study method, to be more precise it is a single-case study approach. As first part of my single-case study I conducted an Analytics Maturity Assessment (provided by TWDI The Data Warehouse Institute, 2015) to find out the current maturity level in the case company and in the second part I analyzed the current data infrastructure to see how analytics are supported in general.

According to Yin (2013, 51) there are five components of a case study research design that are especially important:

1. a case study's questions (what is the current maturity level of analytics)
2. its propositions (future analytics' strategy)
3. its unit(s) of analysis (IT, Controlling, Risk, Other Analytics)
4. the logic linking the data to the propositions (specific maturity level > specific strategy)
5. the criteria for interpreting findings (rival explanations for your findings, N/A)

There are also five rationales that determine whether the single-case study approach is feasible or not. The five rationales are: critical, unusual, common, revelatory or longitudinal case. Selecting a critical case would be critical to your theory or theoretical propositions therefore this case study can be recognized as a critical one because I am comparing the given theoretical framework to this real life case.

As Halper and Stodder state (2014) the TDWI Analytics Maturity Model and assessment tool was created in response to organizations' need to understand how their analytics deployments compare to those of their peers and to provide best-in-class insight and

support. The assessment measures the maturity of analytics across many dimensions that are the key to derive value from analytics. The study tested whether the company is ready for more advanced reporting methods mentioned in the theory or not. The TDWI Analytics Maturity Model consists of five stages: nascent, pre-adoption, early adoption, corporate adoption, and mature/visionary. As organizations move through these stages, they should gain greater value from their investments. The results of this single-case study determine what kind of development strategy regarding the analytics should be applied in the company depending on which one of the above mentioned stages they hold.

3.2 Data Collection

The data collection is an important part of the case study. According to Ghrauri and Grønhaug (2005) one of the best practices to gather such data are in-depth interviews with participants. Secondly distribution of in-depth questions to participants via online tool is another valid method for collecting data (Wilkinson and Birmingham 2003). Wilkinson and Birmingham (2003, 5) stated in their book, that most commonly used research instrument is a questionnaire. Typically, many different ankles are used in surveys, to make sure that the set of answers is what is demanded by the researcher. Most likely, the question types are closed questions, multiple choice or ranking questions, open-ended questions and scale item, Likert-type of questions (Wilkinson and Birmingham 2003, 10.)

Close-ended questions provide all the possible answers to participant, for example, “yes” or “no”. Multiple-choice questions provide also predefined responses, but those should be carefully thought in order to cover all the possible answers. Open-ended questions do not restrict participants’ answers, thus making analyses harder for the researcher. Questions with scaled items provide possibility for asking participants’ opinion about the asked question within predefined list or scale, typically ranging from “strongly disagree” to “strongly agree” (Wilkinson and Birmingham 2003, 11-12.)

I conducted a qualitative survey in order to collect data for the Analytics Maturity Assessment. The survey questions were distributed to 25 contacts via online tool called Webropol 2.0. The questions are provided in appendix 6. The decision about the research instrument for gathering valid information for this study came shortly after analyzing the assessment provided by TDWI because it would have been too time consuming for respondents to get access to the TWDI tool. Better idea was to collect the data personally and then just add the average score as an input to the assessment tool.

In the online tool of this study, the participants were asked to answer only few types of questions, mostly for scale items and multiple-choice. All scaled item questions were using a five-point scale ranging from "strongly disagree" to "strongly agree", see table 3-1 for the full list of scale.

Table 3-1 – Scaled items

Scale	
1	strongly disagree
2	disagree
3	neither disagree nor agree
4	agree
5	strongly agree
6	don't know

Option "I don't know" was also provided, because there was no guarantee that all respondents could be capable of answering every question. Additionally, online tool included several simple open-ended questions to provide the respondents an option to raise their voice for other topics related to the analytics. All of the questions asked from participants were compulsory except for all the open-ended questions.

3.3 Respondents

One of the challenges regarding the case studies is to find correct respondents, key informants, for the topic that is being studied. Reason behind the difficulty in this case is that this research had only limited amount of potential respondents. Another challenge is to motivate respondents and get them respond to the survey. The key group of informants for this study would consist of analysts, controllers and business leaders. The audience was carefully thought and limited to the group that generates and consumes the reports and at the same time they should have some deeper knowledge of the topic.

In order to maximize response rates Wilkinson and Birmingham (2003, 16) recommend to use a short cover letter that explains the purpose of the research. Wilkinson and Birmingham (2003, 16) also recommended to state in this letter if it was anonymous to respond. In addition, reminder emails should be sent to recipients to make sure that they remember to respond. Following this advice, a survey with cover letter and link to the online

tool was created and further sent to participants in middle of April 2015. Participants' cover letter contained the motivation for this research and the potential output for the company.

3.4 Analysis Method

The analysis method used in this research is called "relying on theoretical propositions" as Yin (2013, 36) puts it. Yin (2013, 136) states that in this strategy one should follow the theoretical propositions that led to the case study. Yin (2013, 136) also argues that this strategy of analysis leans on a theoretical framework, propositions and research goals that have initially led to study the current topic. In addition theory and propositions of the research support the structures of a data collection plan

The results can be grouped into five different categories according to the level of progression: nascent, pre-adoption, early adoption, chasm, corporate adoption and mature/visionary (see figure 3-1). Questions may be weighted differently depending on their relative importance. Each dimension has a potential high score of 20 points. Because organizations can be at different levels of maturity in the five dimensions, each section is scored separately and an overall score is delivered as a sum of sections (Halper and Stodder 2014, 17.) Some of the questions aren't scored but rather used for best-practices guidance.

As mentioned earlier the answers are analyzed through TDWI analytics assessment tool which will provide the final maturity level. The tool is sponsored by Cloudera, Tableau and MicroStrategy and therefore the tool is not totally independent but I think it is still valid because the sponsors are among the top advanced analytical tools' vendors.

Figure 3-2 – Analytics stages of maturity (Halper and Stodder 2014)

The nascent stage represents a pre-analytics environment. In this stage, most companies are not utilizing analytics well, except perhaps for spreadsheets. There is no real support for the effort, although there is a critical amount of people throughout the enterprise who are interested in the potential value of analytics and who may be testing analytics

software. Generally, in the nascent stage, the culture is not analytic. In other words, the culture is not data driven and decisions are made based on gut instinct rather than on fact (Halper and Stodder 2014, 10.)

As the company moves out of the nascent stage and into the pre-adoption stage, it is starting to do its homework regarding the analytics. Staff may be reading about the topic and perhaps attending webinars, training or conferences. One or more departments may have invested in some analytics technology such as single instances of a low-cost front-end BI or data discovery tool or a back-end database, data mart, or data warehouse for managed reporting. People are slowly starting to understand the power of analysis for improving decisions and ultimately business outcomes. Some key characteristics of the pre-adoption organization include (Halper and Stodder 2014, 11.)

During the early adoption phase, the company is putting some weight on analytics tools and methodologies. It is thinking about data management and reporting or dashboards. Users often spend a long time moving through the early adoption stage (Halper and Stodder 2014, 12.)

As departments are about to move on from early adoption to corporate adoption and extend the value of analytics to more users and departments, companies must overcome a series of obstacles. This is often why they spend a large amount of time in this phase. There is the obvious challenge of obtaining the right skill set for analysts. There may also be political issues. For example, one department may have been driving the company's analytics effort and brought other departments on board. However, when it comes time to extend the platform or establish more stringent standards and governance, departments begin to fight over who owns the data, or whose particular vision is the best and eventually implemented (Halper and Stodder 2014, 13.)

Corporate adoption is a major milestone in any organization's analytics journey. In corporate adoption phase, end users typically get involved and the analytics transforms how they do business. For instance, users may change how decisions are made by operationalizing analytics in the organization. They will be using different kinds of data, even big data that is semi-structured or unstructured, for their analytics efforts. Organizations that reach this stage of maturity might have constantly addressed certain gaps in organization, infrastructure, data management, analytics, and governance (Halper and Stodder 2014, 15.)

Only a few companies can currently be considered visionary in terms of analytics. At this stage, organizations are executing analytics programs smoothly using a fine tuned infrastructure with well-established program and data governance strategies. Well-governed but flexible data access is available for users so they can explore data and develop visualizations in a self-service fashion and are not completely dependent on IT. Many programs are executed as budgeted and planned initiatives from the company perspective. In the visionary stage, there is a healthy and agile analytics culture that benefits non-traditional users at middle management and even frontline positions (Halper and Stodder 2014, 16.)

Figure 3-3– Analytics Scoring Scale (Halper and Stodder 2014)

SCORE PER DIMENSION	STAGE
4–7.1	Nascent
7.2–10.1	Pre-Adoption
10.2–13.3	Early Adoption
13.4–16.6	Corporate Adoption
16.7–20	Mature/Visionary

Figure 3-4 – Analytics Scoring per Dimension (Halper and Stodder 2014)

DIMENSION	SCORE	STAGE
Organization	10	Pre-Adoption
Infrastructure	7	Nascent
Data Management	11	Early Adoption
Analytics	4	Nascent
Governance	7	Nascent

4 CASE STUDY

This chapter introduces the results of the case study in relation to the theoretical framework and propositions. First, this chapter goes through the survey results. Secondly, the current data infrastructure and reporting methods are analyzed. Finally the results are evaluated against the theoretical framework and propositions and then discussed. Once the results are ready, then the future strategy is proposed to address the original research question.

This case study was performed in a company that will remain anonymous. The company's Senior Management Team (SMT) has identified "fact based decisions" as one of the challenges to focus on towards 2016. In short the idea is that decision making processes should be more based on facts and data analysis and less based on common sense or gut. In order to move with fact based decisions, a research is needed to give some guidance on what is the correct direction.

4.1 Survey Results

The survey was based on questions provided by TDWI. The survey included 35 questions that were categorized by the following topics: organization, infrastructure, data management, analytics and governance. The detailed survey questions can be found from appendix B. The survey took approximately 15 minutes to answer. An online tool called Webropol 2.0 was used to conduct the survey. Respondents had two weeks to respond to the survey during April 2015. An online tool link was distributed straight to 25 contacts individually across the company. The distribution of personal online tool link took place 17.4.2015 and the link was closed 30.4.2015. In total three reminders were sent to respondents to get as many answers as possible and to make sure that all the persons willing to respond to the questions would remember to respond. In total 14 respondents out of 25 finished the survey. The final response rate was 56% which ended up being lower than expected.

Based on the survey answers that were imported to the analytics maturity assessment tool, it gave the case company an overall score of 10. The maximum score is 20. The score of 10 points gives the company a pre-adoption rating. The rating scale was nascent, pre-adoption, early adoption, corporate adoption and mature. This means that the company is at the moment taking steps to correct direction but is not on mature or advanced level yet. The

score is a little lower but still in line with the respondents' expectations that were evaluated by a set of preliminary questions. The most common answer related to expectations suggested that the company's analytics competence and activity is on average level compared to competitors.

If we compare the results by categorizing them under organization, infrastructure, data management, analytics and governance layers, we can find some more insights on how the overall score is generated (see table 4-1). The highest rating was scored from analytics (12) and organization (11.5) and lowest score from governance (9.75) and infrastructure (10). Data Management (10.5) scored between those two groups. The rating maximum and minimum score difference was only 2.25 points between the five categories. That is rather low, given that the maximum amount of points was 20. Analytics and Organization scored high, because respondents feel based on the answers that data-driven approach and culture is supported in the company. Most of the respondents felt like there were appropriate tools and know-how in place for 51-75% of analysts and data scientists. This probably decreased the score of analytics because it should be near 100%. Even though analytics are not delivered in automatic way, there is a strong belief that company knows which questions are important and that correct questions are answered by manual analytics. When the analytics are serving the need of managers or business it is a good sign. However this might lead to situation where nothing new is discovered because the current way of doing things is working just fine. There should be some sort of pressure for the analysts to find something new in order to gain real benefits and competitive advantage. There is actually no point of taking advanced analytics into use, if there is no one pushing the analytical efforts to find some new insights.

Data management, governance and infrastructure got the lowest score. There are several potential reasons for this. First, I could imagine that whilst the data warehouse has been built it was not created using Guo et al. (2006) theory of triple-driven data modelling methodology. If the DW was created by using Kimball's (1997) original theory there might have been lack of user involvement. It is possible that the data-driven approach has been used or then the people involved are not working for the company anymore. Thus, analysts are not too familiar with the concept. Even if analytics are seen as an important part of the business there is a lack of data governance and overall understanding amongst the analysts. This means that there is locally no person or committee with a role that would be fully committed to governing the data. It might seem from analyst' perspective that the data is fragmented all over the company and it is more like Wild West than sophisticated information-driven

company. Data standards and quality are not looked after carefully and everyone is playing by their own rules when interpreting the data. The many "don't know" answers just emphasize the limited leadership in data governance. In addition six hits for "We haven't had time for data management and ownership policies but we know we need to do that" and seven hits for "we don't have analytics governance team in place" are also signs that are demonstrating the poor data management. This is something that should be considered when the new strategy is designed and implemented.

If the company's score is compared against industry and corporate size separately then the case company seems to be lacking behind the average performance. But when the score is compared against industry and corporate size together then the case company is ahead of its competitors. That is good sign because it means that similar competitor companies by the size and industry are doing worse on average than the case company. On the other hand in each category there are top notch companies that are performing plus five points better than this company. To be honest, the company is not the worst but it still has a long road ahead if the goal is to reach the top performers.

Overall the survey produced controversial answers. There is no clear path for analysts to follow when it comes to data management, governance and infrastructure. The infrastructure is slightly lacking behind compared to market average since the business data warehouse has just recently been released and the company is taking the first steps on right path. The immaturity shows in poor governance and the environment is not well organized. Despite of the poor governance analytics scored 12 points which is way above the industry average score. This could mean that the analysts are good, but their efforts are not supported by data leadership. This creates a problem of inefficient reporting which can cumulate to hundreds of hours of unnecessary work in only one fiscal year. According to SAS' survey (2015, 19) in Nordic 63% of companies in Finland admit that they need to upgrade their current data center infrastructure and 90% think more and new data would give them a competitive advantage. Based on these results there is going to be major investments to infrastructure and effort to be taken in order to improve the data governance activities. Therefore the case company should be ready to invest in BI infrastructure to keep up with the competition.

Table 4-1 - Scores of Analytics Maturity Assessment by Categories

Overall score	**10,0/20**
Organization	11,50
Infrastructure	10,00
Data Management	10,50
Analytics	12,00
Governance	9,75

Organization 11,5/20	Min	Avg	Max
Industry	0	11,93	18,13
Corp Size	0	11,64	18,5
Industry/Size	5,5	9,82	13
Overall	0	11,49	20

Infrastructure 10,0/20	Min	Avg	Max
Industry	0	10,08	15
Corp Size	0	10,33	17
Industry/Size	5	9,05	10,5
Overall	0	9,97	18

Data Management 10,5/20	Min	Avg	Max
Industry	0	9,71	14,5
Corp Size	0	10,09	17
Industry/Size	7,5	9	10,5
Overall	0	9,74	17,5

Analytics 12/20	Min	Avg	Max
Industry	0	10,2	16,25
Corp Size	0	10,03	18,25
Industry/Size	6,25	9,2	12
Overall	0	9,94	19,25

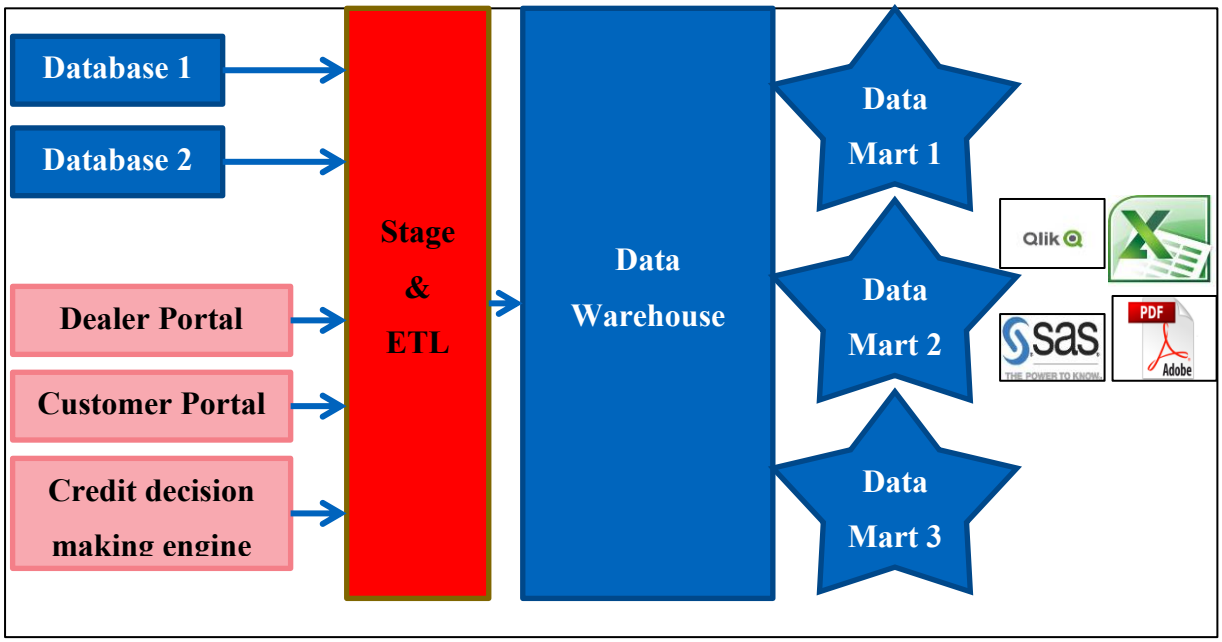
Governance 9,75/20	Min	Avg	Max
Industry	0	9,79	16,5
Corp Size	0	9,56	18,5
Industry/Size	4,75	8,98	11,25
Overall	0	9,42	18,75

4.2 Analysis of BI Architecture and Reporting Methods

Currently in the case company there is a data warehouse in place. The architecture follows exactly the formula of classic data warehouse which was discussed earlier in chapter two. Only the data marts are still under development. There are two main databases and in addition there are credit decision making engine, dealer and customer portals that use external data. The portals generate for example clickstream data. There are plans for several data marts to be implemented for different departments. Data marts are not yet in use, but

there are projects going on to develop those. In addition there are several applications to manipulate and access the data. Reporting standards include excel and pdf files which are shared in management information systems. Figure 4-1 illustrates the planned architecture and analytics environment in the case company.

Figure 4-1 – Data Warehouse architecture and Business Intelligence environment



The results of BI architecture and reporting methods analysis are in line with the survey results. The company is somewhere near the average level or little behind. The reason why company has only recently developed its data warehouse is related to the past. There have been several efforts made to build the DW but the company has not succeeded until recently. Another complete research could be done about this topic but I am not going to details in this study. However, as was mentioned in the theory, there are many pitfalls regarding the implementation of DW. Probably the company was not well prepared to the large amount of work that was required from the whole organization. Therefore it is not so surprising that the company did not succeed on the first try. However at the moment the company's data-driven culture is a positive driver that allows the analytics to develop. Before the data marts are in place and Qlik software in full use the company could evaluate whether to use Qlik or Tableau. According to Sallam et al. (2015) Tableau is at the moment a little ahead of Qlik. If this gap grows bigger in future it might be worth of reconsidering the tool that is going to be used.

As mentioned in survey results part the lack of data governance is slowing down the development and creating friction in reporting processes. It is clear that the company is

adjusting to this new data-driven culture and everything is not in order yet, only the most critical functions are running flawlessly that allow the company make the most important decisions. However there have not been great efforts taken to support the valuable work of skillful analysts. Most of the analysts are capable of using SQL queries and SAS and even data mining methodologies. Unfortunately large portion of the time goes to the data collection process. This time is away from analyzing the data. After all analyzing the data is the most beneficial for the company if it seeks to discover and unlock new value for the business.

4.3 Proposed Strategy for Future Analytics

According to the survey results and my architecture analysis the case company as a whole is not currently mature enough to rush into advanced analytics or big data. Even though some of the respondents seem to be qualified and ready for the advanced level the company's journey to become truly data-driven is still waiting for takeoff in the departure lounge because the fundamental infrastructure and leadership in data management are lacking behind. Based on everything I have read, researched and analyzed I would suggest six strategic moves that the company should take. The reasons why these steps are necessary will be explained below.

1. *Analyze and renovate the foundations of data management*
2. *Establish and promote data governance committee*
3. *Communicate the data management and architecture principles to every analyst in order to increase the level of unity and engagement regarding the BI&A*
4. *Incrementally develop a modern data warehouse architecture that would include Hadoop, Analytics Mart and Predictive Analytics solutions*
5. *Reconsider between Qlik and Tableau*
6. *Conduct researches also in future to keep up with the development*

First the company should start with analyzing and renovating the current foundations of data management because there is no clear leadership in data management. The lack appears as a minor chaos amongst the analysts and there are no common guidelines on how to manipulate the data. As the survey revealed there is no full understanding of the architecture amongst the analysts and leaders. There is no need for big transformation but instead the company should just recap the data management structure and do its homework based on that. Ideal group for recap and brainstorming would consist of senior management

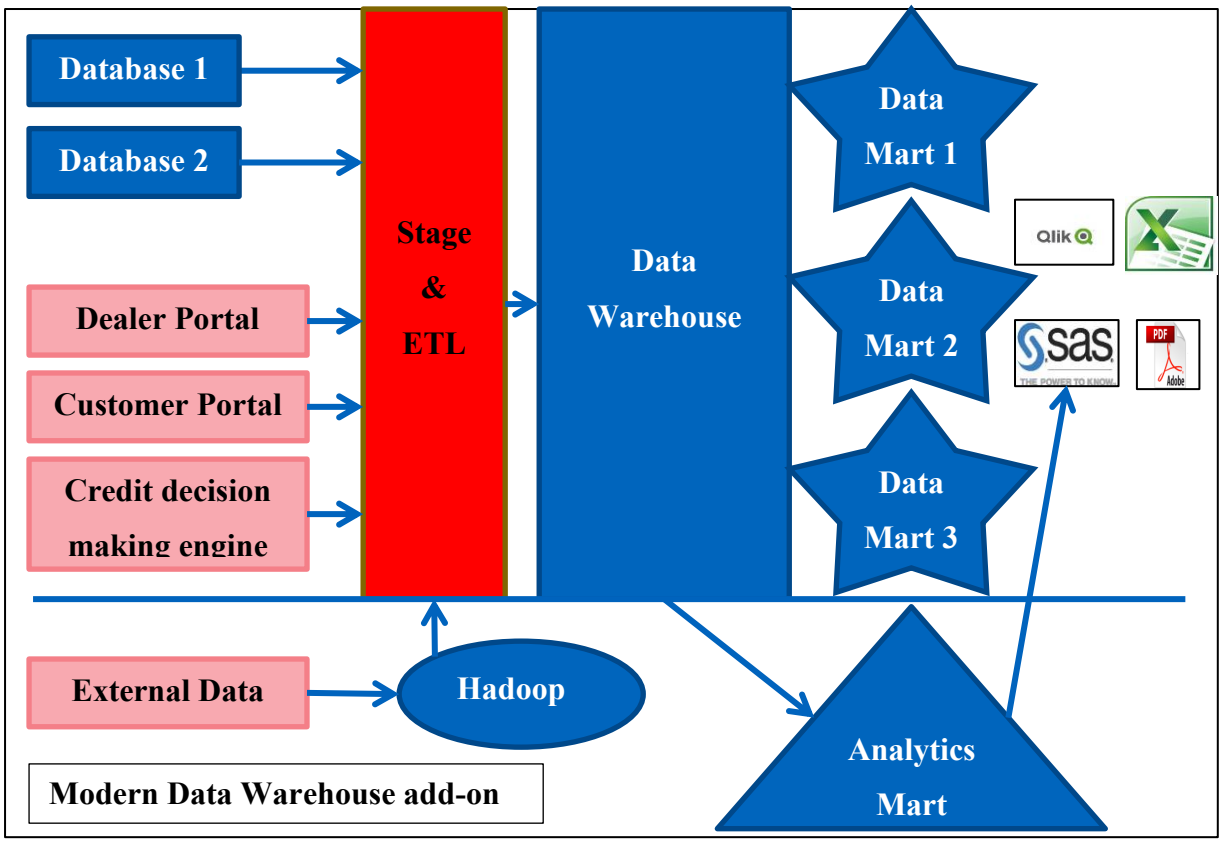
team, IT and business analyst representatives. This would ensure that all the important stakeholders are heard. Even if the infrastructure was good enough I would suggest every company to have sanity check on its processes every now and then to ensure continuous development. Therefore I think this first step is reasonable enough to be included as a starter for the future strategy.

Next, establish a data governance committee. The company needs a data governance committee because of the above mentioned lack of common guidelines. This committee would consist of analysts from each department. These individuals would lead the governance actions as a team. This team should actively arrange workshops to all analysts to improve the infrastructure knowledge and to create team spirit. Once there is knowledge and team spirit it is easier to engage and share best practices among the group. Governance committee actions could also include such policies and processes that aim to collect, maintain and update for example data definitions and queries that are held in query library. Currently there is no common process for creating and updating queries and it relies on individual analyst's memory. However there are hundreds of queries and the current approach with no control is error-prone. Someone could argue if this is a necessary step because it could limit the freedom of doing things as one likes. The point is not to limit analysts but to create common guidelines on how to work on data. Finding the balance between freedom and chaos is important. It would be then easier for example new employees to find out how the analysts are working in the company and for current analysts to know how their colleagues work. Once you know how your colleagues work it is easier to share best practices and make the analyzing more efficient.

Third, training and updating the analysts' knowledge regarding current infrastructure, data management and governance should be conducted. Responsibilities and roles of data management should be clearly defined. Defining the content, structure meaning and usage of data should be updated and taught to current as well as new employees that enter the company. For example if an analyst is recruited there should be a defined data understanding path in place. The data is in core of an analyst's task and there is no room for misunderstanding. For example basic guidance of how the data is treated, for what purposes and by who could do the trick to avoid misinterpretation. Unity and engagement are important and would generate discussion on what methods to use. For example whether to use Qlik or Tableau would require further analysis than just yes or no approach.

Fourth would be the actual implementation of modern data warehouse architecture in order to gain more value from external data sources. This could be done once the classic data warehouse architecture is fully established and assimilated. I would suggest incremental development instead of radical one because incremental approach would suit the case company that already has a data warehouse. Incremental approach would also support the adaption of a new culture and it would hold only a moderate risk. The current architecture would be suitable for a Hadoop + analytics mart combo that would drive external data analysis and predictive analytics. The option of using external consultants in this augmentation process should not be excluded, because those might bring beneficial fresh ideas to the company. According to SAS survey (2015, 21) 15% of companies already use or are implementing Hadoop system in Finland. This means that there are already potential competitors few steps ahead. This observation should at the latest encourage the use of Hadoop. On the other hand literature (Gartner 2015) reveals that many companies are struggling to find employees with enough big data or Hadoop competence. Therefore it would be realistic to say that the development should be done within three years instead of e.g. one year. Luckily, at least Aalto University in Helsinki is including some Hadoop modules in data science courses to fill this gap.

Figure 4-2 – Modern Data Warehouse Architecture for the Case Company



Even though it seems reasonable to advance to data warehouse and Hadoop combination in this case it is recommended to stress some critique on this choice. There is no real knowledge or guarantee whether this investment on advanced analytics is profitable in a long run or not. I would compare this kind of experimental or exploration to gold mining. There is certain probability that the gold exists but there is really no answer how much gold there is and if it covers the mining costs. Same goes with the data. Companies cannot be sure that the data mining is profitable, only the time will tell. However, the risk that the competitor takes the gold instead of you exists. The question is if this company wants to take that risk or not. I would suggest to avoid this kind of risk and therefore I argue that the company should apply advanced analytics in future even though there is no guarantee of return. This kind of exploration would go in hand with previous experiments on other innovations in this company. The company culture supports this kind of experiments and therefore it should not be a barrier in the implementation of Hadoop technology.

Fifth step would be the comparison between Qlik and Tableau. Tableau seems to be ahead of Qlik according to the research conducted by Gartner (2015). Since the company is still in the pre-adoption phase it would not be too late to reconsider if Tableau is actually better in the long run. Usually switching the software when it is widely used in a company will be very time-consuming task. I would suggest that in the beginning there could be few licenses for Tableau just to make sure that company picks the correct vendor. Even if the Tableau is not feasible or worth of the higher price the company could continue using Qlik as planned initially.

Finally, because the field of BI&A is constantly changing and developing, the company should conduct more researches in future via help of universities, like this thesis. Offering master's thesis options for students would bring the enterprise the newest information of potential future solutions. Not only the company can benefit of research but also the local community is learning if it is offered more chances to conduct researches. If the company wants to be on top of competition it usually needs first mover advantage and that advantage is gained by early research. Of course another option is to choose the follower strategy. If the company does not have enough resources or thinks that early bird pays off it can follow and benchmark the first mover companies. Personally I think that in this field of data mining, it is hard to benchmark the best practices because of the privacy policies. For that reason I believe that in this field it pays off to conduct early research.

5 SUMMARY AND CONCLUSIONS

The final chapter of this research consists of research summary, limitations and suggestions for future research. The chapter shortly addresses answers to the original research questions. Furthermore it will recap the results of the case study and present my general thoughts on this topic.

5.1 Answers to Research Questions

Here are short answers to the original research questions:

1. *What does the state of art modern BI&A solutions look like today?*

Compared to the traditional and classic data warehousing solutions, (Turban et al. 2011) modern data warehousing takes the external and unstructured data forms and types into account in a more detailed level. Also, there are no clear steps in earlier data modelling theories (Moody and Kortink 2000, Guo et al. 2006) on how the data should be modelled in case of more complex data types and sources beyond the data warehouse solution.

I would argue that today the state of art solution would include a data warehouse solution combined with a Hadoop technology. This approach is strongly supported by D' Antoni and Lopez (2014). There were several modern solutions provided in literature ranging from Hadoop only to in-memory computing on how the data flow should be treated. Eventually it seems that the combination of data warehouse and Hadoop is the most appropriate approach for many companies because their designs support each other and many companies already use the classic data warehouse solution. The modern theories regarding this combination are also able to address the challenges regarding complex data types and their data models. Because the combination is trying to unlock value from internal and external data as well as structured and unstructured data it outperforms the earlier solutions that are only processing structured and internal data efficiently. The earlier approaches are simply not capable of taking the advantage of the new types and volumes of data efficiently.

In order to cope with the trend of increasing data volumes companies need to invest more and more in costly additional horsepower to their existing EDW solution every year (Dull 2014) in the case of traditional data warehouse solution. To avoid this, companies should consider alternative ways to manage their data. My proposed alternative solution makes more sense cost-wise and efficient-wise in a long rung as well due to low cost of

Hadoop storage and high capacity of Hadoop processing power. Because of the low cost of Hadoop storage, you could store both versions of the data in the HDFS: the non-cleansed data and the after transformed data. All of the data would be in one place, making it easier to manage, reprocess and analyze at a later date. As Dull (2014) puts it processing data in Hadoop frees up EDW resources and gets data processed and transformed into your EDW quicker so that the analysis work can begin.

The modern BI&A is aiming to tame predictive analytics and big data with the use and help of existing data warehouse methodologies. Thus, companies are able to more efficiently to advance to the next level instead of "starting from a scratch"-solutions. The modern theory proves that the data warehouse and Hadoop combination's cost-efficiency and elasticity of handling different kinds, types and volumes of data is well ahead of the earlier solutions. Based on these facts the combination of data warehouse and Hadoop is currently the state of art solution for modern BI&A. The modern data warehouse add-on illustrated in figures 2-12, 2-13 and 4-2 shows the new addition compared to traditional data warehouse approach illustrated in figure 2-5. It will be part of the staging and ETL process in addition to predictive analytics module in the actual analyzing process. Predictive analytics is totally missing in earlier theories and this would be significant addition to the traditional data warehousing solution.

2. *What is the current BI&A maturity level of the case company?*

According to analytics survey provided by TDWI and architecture analysis the current overall maturity level in the case company was close to average compared to industry. In other words the company was on pre-adoption level, given that the range was nascent, pre-adoption, early adoption, corporate adoption and mature. The company scored 10 out of 20 points. There were five categories that were measured (data management, governance, infrastructure, analytics and organization). The company scored the highest points on analytics and organization, but lower scores on governance and infrastructure, data management was rated between these two edges. The architecture analysis was in line with the analytics survey and provided similar results that the case company was on average level.

However I think that it is advisable to acknowledge that there is not a lot of theories on how to measure BI&A maturity levels. TDWI is providing only one method and I would say that relying on only one method is quite vulnerable to errors. Thus, I feel that the results of the survey are not bulletproof but on the other hand the difference marginal is probably not

significant. The company could be only slightly more or less advanced. Gladly, the architecture analysis is balancing the results and giving another perspective, which makes the overall maturity score more reliable. Also because of the fast development in the business environment the maturity level could be on different level compared to other companies given that the time and date was something else that it was in this study.

3. *What kind of future BI&A-strategy would be optimal based on the literature review, survey results and architecture analysis?*

Because the company was not mature enough to rush in for advanced analytics I would propose incremental strategy instead of radical strategy for the case company. As was mentioned in theory (Weir 2002) for data warehouse solutions incremental approach is always better, because the system is company-wide and complex. There were many positive signs of data-driven culture, but due to fact that company is in developing phase the only appropriate option would be first reviewing the data management principles and move on from there. Second step includes the data governance committee. As Niemi (2014) stressed the importance of data governance I cannot see a strategy without this step. The third steps would be resulting from the second step once the committee is in place and ready to take actions. Fourth, the combination of data warehouse and Hadoop is the solution that the case company should aim at in the next three years. The goal is to mine the gold incrementally. Therefore three years would be optimal before it is too late. On my own experience the trends of specific methods lasts only a short time but also there is no point of implementing something that you are not ready for. Fifth would be the comparison between Qlik and Tableau as explained in section 2.8.1 and 4.3. Finally future research is needed if the company chooses first-mover strategy. Below is included the full six step strategy.

1. *Analyze and renovate the foundations of data management*
2. *Establish and promote data governance committee*
3. *Communicate the data management and architecture principles to every analyst in order to increase the level of unity and engagement regarding the BI&A*
4. *Incrementally develop a modern data warehouse architecture that would include Hadoop, Analytics Mart and Predictive Analytics solutions*
5. *Reconsider between Qlik and Tableau*
6. *Conduct researches also in future to keep up with the development*

There is not much literature on Hadoop adoption strategies or advanced analytics adoption strategies except for the technology adoption in general (e.g. Saarinen 2013). Therefore I am following Weir's (2002) first advice when it comes to the adoption: "The project must fit with corporate strategy and business objectives". I tried to design the strategy steps so that they would fit the company culture, strategy and business objectives. I think these implementation steps are quite straightforward and doable instead of fancy plans that never meet the expectations.

5.2 Research Limitations

Although the research has achieved its goal to answer to the research questions that were designed in the beginning of the study, it still has its limitations. Firstly, the sample size in this paper was low. Secondly there was lack of alternative maturity analyzing methods, making the analysis of survey quite narrow. Thirdly, the time was limited and the scope was restricted due to fact that this was master's thesis, a common exercise to identify the researchers with highest potential.

5.3 Suggestions for Future Research

There are a numerous of good future research possibilities in modern BI&A research. I think that research of the actual implementation of Hadoop Technology is needed. Also measuring the analytics maturity needs an alternative way alongside the one provided by TDWI. TDWI model was sponsored by few top advanced analytics vendors. Therefore an independent study by a university level research would be a great addition to this field.

5.4 Final words

I can honestly say that this research was both a painstaking and an enjoyable process. At times it was hard to decide what parts of the literature are necessary and essential. Due to the complex nature of business intelligence and analytics it may be hard for the reader to catch the idea if there is not enough or if there is too much information. I left out a lot of information but I still hope that there is not too much of everything. I could have focused more specifically on some of the topics I discussed but then again I think it is more important to understand the concept from wider perspective instead of focusing on only one thing.

The most enjoyable moments for me during this research were the ones when I was able to outline the big picture. There are many variables in this equation, which are constantly

developed. In one year there might be so much new stuff available that if you are not actively following the scene you may easily drop out of the development. I am also grateful that I was able to draw a quite straightforward strategy for the case. In my opinion the maturity test provided valuable information for the company because sometimes it can be hard to tell "where we stand at the moment". An assessment provided by external research party is probably the best method to get a neutral answer to that question.

All in all I think this was the kind of study that was beneficial for both the researcher and the company. There were something new, something old and something that could be improved. As Albert Einstein has said "You have to learn the rules of the game. And then you have to play better than anyone else". Now that the company knows the rules of modern BI&A game they have to only play their cards better than anyone else to ensure their success in future.

REFERENCES

Adelman S. and Moss L., 2001. Data Warehouse Risks, Journal of Data Warehousing, Vol 6, No. 1.

Bantleman John 2012. The Big Cost of Big data, Forbes, http://www.forbes.com/sites/ciocentral/2012/04/16/the-big-cost-of-big-data/ Retrieved 04-03-2015

Barton Dominic and Court David, 2012. Making advanced analytics work for you, Harvard Business Review, Vol 90 No. 12 79-83.

Chen Hsinchun, Chiang Roger H.L. and Storey Veda C. 2012. Business Intelligence and Analytics: From Big data to Big Impact, MIS Quartely, Vol. 36 No. 4, 1165-1188.

Clemons, E.K., 1995, Using Scenario Analysis to Manage the Risks of Reengineering, Sloan Management Review, Iss. 36, pp. 61-71.

Codd, E.F. 1970. A Relational Model of Data for Large Shared Data Banks, Communications of the ACM, 13 (6), June: 377-387.

Cooper, R., Markus, M., 1995. U: Human Reengineering, Sloan Management Review, Issue 36, pp. 39-50.

D'Antoni Joseph and Lopez Karen 2014. The Modern Data Warehouse – How Big data Impacts Analytics Architecture, Business Intelligence Journal, Vol. 19 No. 3, 8.

Davenport Thomas H., Barth Paul and Bean Randy 2012, How 'Big data' is different, MIT Sloan Management Review, Vol. 54 No. 1 43-46.

Dull Tamara, 2014. A Non-Geek's Big data Playbook – Hadoop and the Enterprise Data Warehouse. SAS [http://tamaradull.com/2014/02/17/white-paper-a-non-geeks-big-data-playbook-hadoop-and-the-enterprise-data-warehouse/] Retrieved 22-03-2015

DWHworld.com, 2015. Simple Star Schema Model for Sales Fact [http://www.dwhworld.com/2010/10/simple-star-schema-model-for-sales-fact] Retrieved 19-05-2015

Ettlie, J.E., Bridges, W.P., O'Keefe, R.D., 1984, Organization Strategy and Structural Differences for Radical Versus Incremental Innovation, Management Science, Iss. 30, pp. 682-695.

Flick U., 2007. Designing Qualitative Research, Sage Research Methods Online.

Gartner, 2015. "Survey Analysis: Hadoop Adoption Drivers and Challenges. http://www.gartner.com/document/3051617 [Retrieved 19-05-2015]

Gersick, C.J.G., 1991. Revolutionary Change Theories: A Multilevel Exploration of the Punctuated Equilibrium Paradigm, Academy of Management Review, Iss. 16, pp. 10-36.

Ghrauri, P., Grønhaug, K., 2005, Research Methods in Business Study, A Practical Guide, Harlow: Prentice Hall, Europe.

Guevara Jamie, Hall Linda, Steggman Eric, 2011. "IT Key Metrics Data 2012: Key Infrastructure Measures: Storage Analysis: Current Year", Gartner.

Guo Yohong, Tang Shiwei, Tong Yunhai & Yang Dongqing 2006. Triple-Driven Data Modelling Methodology in Data Warehousing: A Case Study, Association for Computing Machinery (ACM) DOLAP '06 Proceedings of the 9th ACM international workshop on Data warehousing and OLAP, P 59-66.

Halper Fern and Stodder David, 2014. TDWI Analytics Maturity Model, TDWI Research.

Helenius Jere, 2014. Business Intelligence & Data Modelling presentation, Aalto University School of Business, Business Intelligence, Guest lecture slides by Siili Solutions, September 19th, 2014.

Hwang M. and Xu H., 2005. A Survey of Data Warehousing Success Issues, Business Intelligence Journal, Vol 10. No. 4.

IBM Big data at the Speed of Business: Big data use cases, 2015. http://www-01.ibm.com/software/data/bigdata/use-cases/data-warehouse.html Retrieved 04-03-2015

Jain Scweta and Nandi Sujay, 2014. Data warehouse augmentation, Part 2: Use big data technologies as a landing zone for source data, IBM developerWorks.

Janssen Dale and Janssen Cory, 2015. Definition - What does Enterprise Data Warehouse mean, Techopedia Janalta Interactive Inc. (http://www.techopedia.com/definition/26204/enterprise-data-warehouse) Retrieved 17-01-2015.

Johnson Jeanne E., 2012. Big data + Big Analytics = Big Opportunity, FinancialExecutive, 51-53.

Jarvenpaa, S., Stoddard, D, 1998, Business Process Redesign: Radical and Evolutionary Change, Journal of Business Research, Vol. 41, pp. 15-27.

Kaldeich Claus and Oliveira e Sá Jorge, 2004. Data Warehouse Methodology: A Process Driven Approach, Advanced Information Systems Engineering, Universidade do Minho, 1-16.

Kelley Moe, 2009. Answer Any Question With Fact-Based Decision Making, The CEO Refresher. (http://www.refresher.com/Archives/amokfact.html) Retrieved 01-02-2015.

Kendle Noreen, 2005. The Enterprise Data Model, *Published in TDAN.com.*

Kimball, R. 1996. The Data Warehouse Toolkit, New York: J. Wiley & Sons.

Kimball, R. 1997. A Dimensional Manifesto, DBMS Online, August, 1997.

Kulin Samu, 2013. Bachelor's Thesis: Key Challenges That Managers Face with Big data, Aalto School of Business, 1-29.

Laney Douglas 2011. "3D Data Management: Controlling Data Volume, Velocity and Variety". META Group. Available (http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-DataManagement-Controlling-Data-Volume-Velocity-and-Variety.pdf) Retrieved 17-01-2015.

McAfee Andrew and Brynjolfsson Erik, 2012. Big data: The management revolution, Harvard Business Review, Vol. 90 No. 10, 59-68.

McKinsey Global Institute (MGI) 2011. "Big data: The next frontier for innovation, competition and productivity." MGI Report.

Moody Daniel L & Kortink Mark A.R., 2000. From Enterprise Models to Dimensional Models: A Methodology for Data Warehouse and Data Mart Design, Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2000), 1-12.

Nadler, D.A., Shaw, R.B., Walton, A.E., 1995, Discontinuous Change, Jossey-Bass Publishers, San Francisco, CA.

Nguyen, T.M., Min Tjoa, A., and Trujillo, J., 2005. Data warehousing and knowledge discovery: A chronological view of research challenges, Institute of Software Technology and Interactive Systems, Vienna University of Technology.

Niemi Erkka, 2014. Data Governance & Management presentation, Aalto University School of Business, Business Intelligence, Guest lecture slides by Siili Solutions, September 12th, 2014.

Oracle, 2000. Oracle8*i* Data Warehousing Guide Release 2 (8.1.6) Part Number A76994-01 (http://docs.oracle.com/cd/A87860_01/doc/server.817/a76994/schemas.htm) Retrieved 20-02-2015

Pohjola Matti, 2010. Taloustieteen oppikirja, WSOYpro 1-264.

Predictive Analytics Today, 2014. Top 15 Extract, Transform, and Load, ETL Software (http://www.predictiveanalyticstoday.com/top-free-extract-transform-load-etl-software/#content-anchor) Retrieved 26-01-2015.

Reeves L, 2009. Manager's Guide to Data Warehousing. Hoboken, NJ, Wiley.

Ruponen Jukka, 2012. Big data and Business Analytics presentation, Aalto University School of Business, Management Information Systems, Guest lecture slides by IBM, September 19th, 2012.

Ruponen Jukka, 2014. IBM Business Analytics: Driving Business Optimization with Trusted Data, Aalto University School of Business, Business Intelligence, Guest lecture series by IBM, September, 2014.

Saarinen Timo, 2013. Strategies for successful BPR implementation, Aalto University School of Business, ICT Enabled Business Process Development, lecture series in fall 2013.

Sallam Rita L., Hostmann Bill, Schlegel Kurt, Tapadinhas Joao, Parenteau Josh and Oestreich Thomas W. 2015. "Magic Quadrant for Business Intelligence and Analytics Platforms", Gartner, http://www.gartner.com/technology/reprints.do?id=1-2AD1WP8&ct=150223 (Retrieved 04-03-2015)

Sarsfield Steve, 2009. The Data Governance Imperative: A business strategy for corporate data", IT Governance Publishing, 1-67.

SAS Institute, 2015. The 2015 Nordic survey on Big data and Hadoop, SAS Institute.

Saunders, 2009. Cooking Up a Data Warehouse, Business Intelligence Journal, Vol. 14, no. 2.

Solomon M, 2005. Ensuring a Successful Data Warehouse Initiative, Information Systems Management Journal.

Sue Valerie M., Ritter Lois A 2012. Conducting Online Surveys 2nd Edition, Sage publications

The Data Warehouse Institute (TWDI) 2015. (http://tdwi.org/pages/research/maturity-models-and-assessments.aspx) (Retrieved 22-03-2015)

Turban Efraim, Leidner D., McLean E., and Wetherbe J. 2006. Information Technology for Management 5th ed., New York, Wiley.

Turban Efraim, Sharda Ramesh and Delen Dursun, 2011. Decision Support and Business Intelligence Systems 9th Edition, Pearson Education Inc., Publishing as Precentice Hall 326-373.

UC Berkeley's School of Information Management and Systems, Executive Summary How Much Information? 2003. (http://www2.sims.berkeley.edu/research/projects/how-muchinfo-2003/printable_report.pdf) Retrieved 17-01-2015.

Yaddow Wayne, 2014 Meeting the Fundamental Challenges of Data Warehouse Testing, Business Interlligence Journal, Vol. 19 No. 3, 28-34.

Weir, R., Peng, T., and Kerridge, J., 2003 Best practice for implementing a data warehouse: A review for strategic alignment. In Proc. DMDW.

Wixom B. and Watson H. 2001. An Empirical Investigation of the Factors Affecting Data Warehouse Warehousing Success, MIS Quarterly, Vol. 25 No. 1.

Wilkinson, D., Birmingham, P., 2003, Using Research Instruments, A Guide for Researchers, Published in the Taylor & Francis e-Library, pp. 5-41

Williams Steve, 2014. Big data Strategy Approaches: Business-Driven or Discovery-Based, Business Intelligence Journal, Vol. 19 No.4, 9-15.

Yin, R.K. (2013) Case Study Research: Design and Methods 5th edition, Sage Publications.

Appendix A: Illustrations of Data Models

Figure A1 – Conceptual Data Model

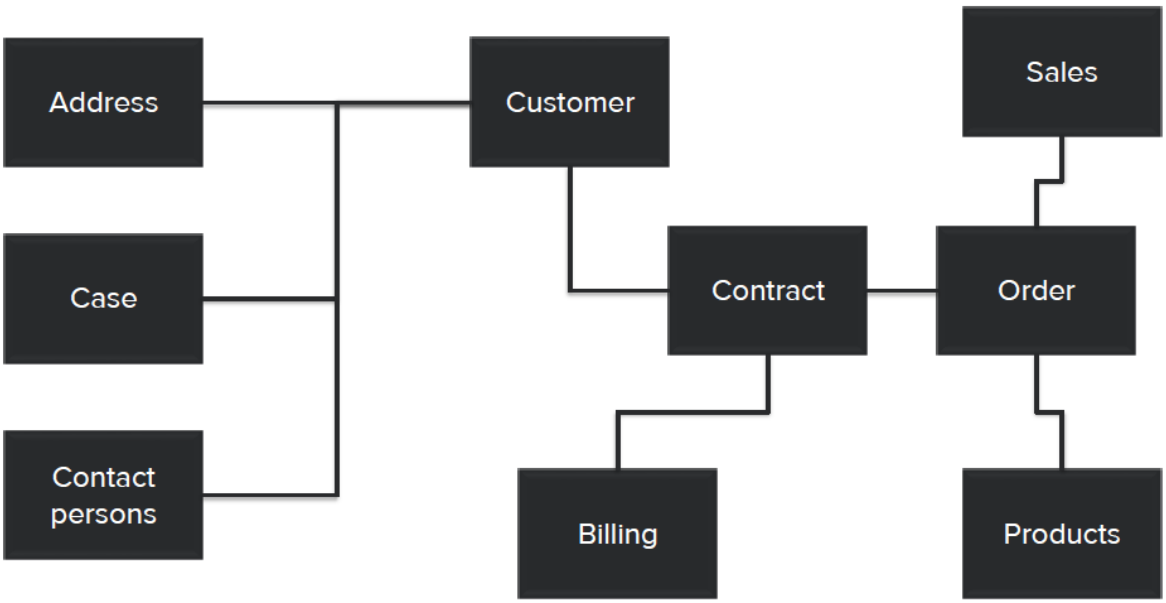


Figure A2 – Logical Data Model

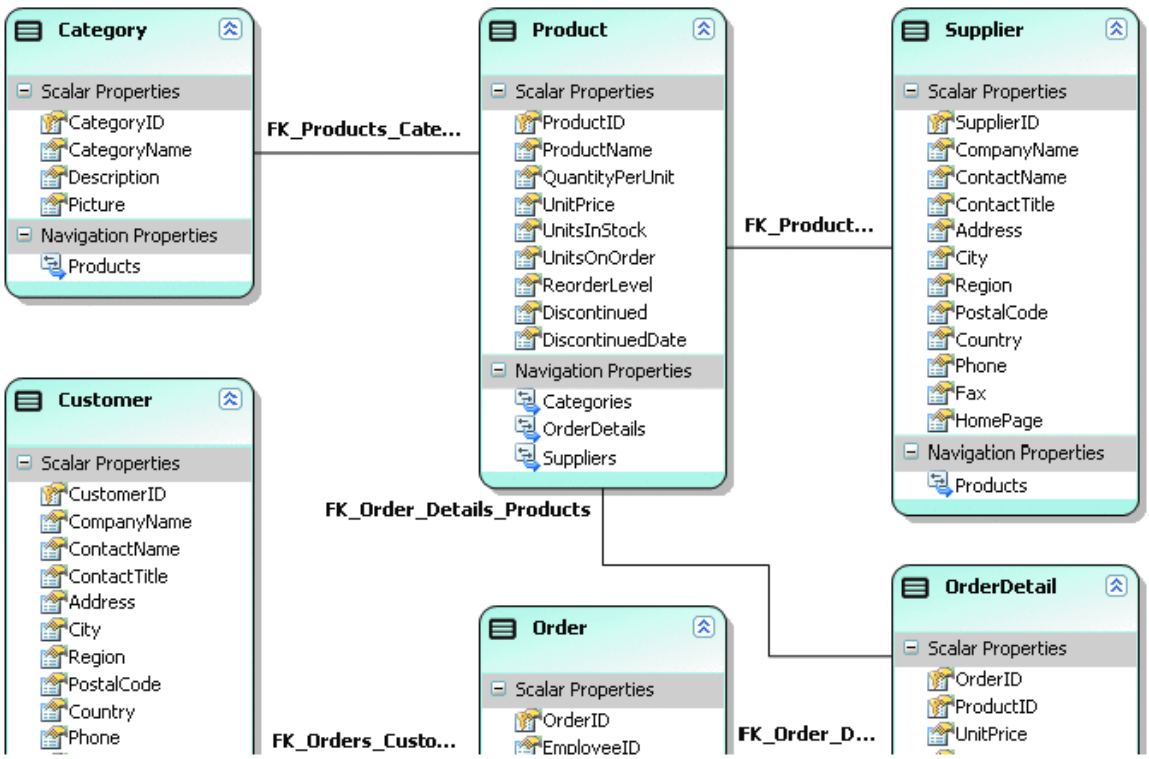


Figure A3 – Physical Data Model

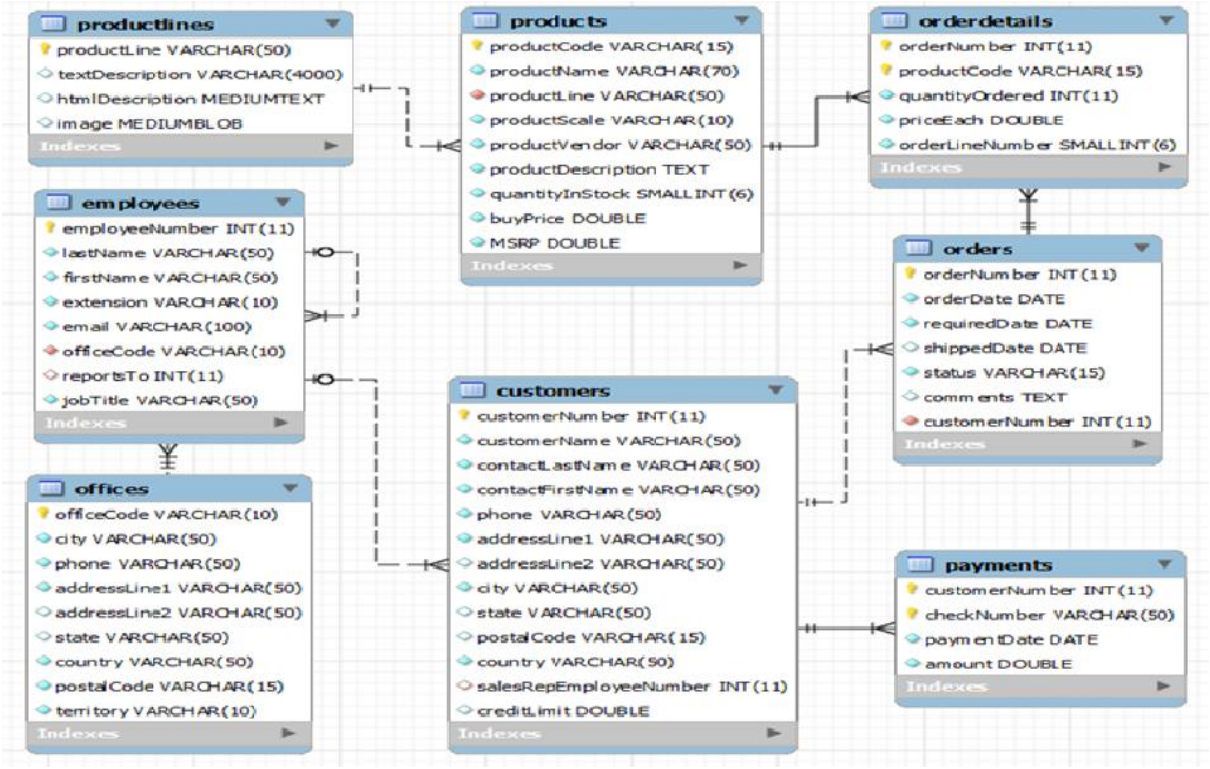
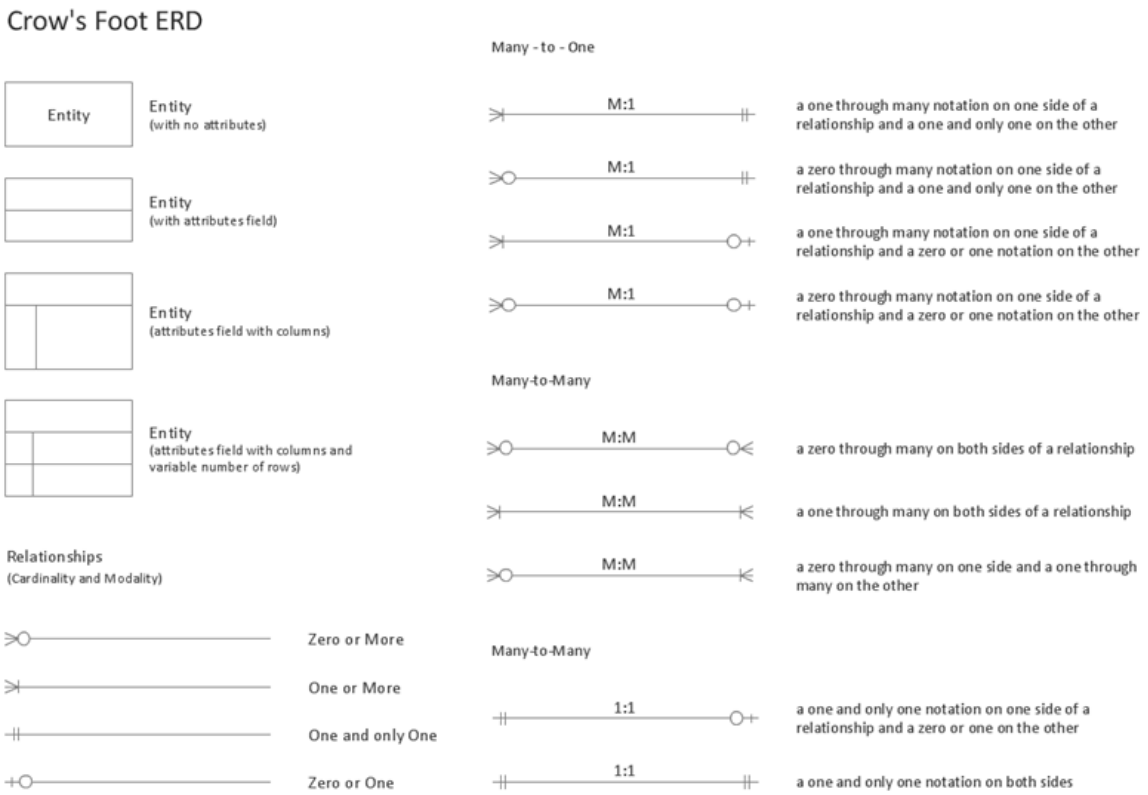


Figure A4 – Icons used in Data Model graphs



Appendix B: Survey Questions

TDWI Analytics Maturity Model Assessment (modified for the case)

Thank you in advance for participating in this benchmark study on analytics. TDWI's goal is to help organizations learn from peers to gain new business advantages from analytics.

BACKGROUND: This survey asks questions about your organization's current strategies for analytics. Through participation in this survey, you will be able to benchmark where you are in your analytics journey relative to your peers. This can help you more effectively plan for the future.

PURPOSE: This 10-15 minute survey asks a series of questions across five dimensions related to analytics. These are Organization, Infrastructure, Data Management, Analytics, and Governance. At the end of the survey you will receive your score in each of these dimensions relative to your peers. We ask that you provide an honest appraisal of your analytics progress to ensure that you and others taking the benchmark survey receive the best possible insight.

WHO SHOULD TAKE THIS ASSESSMENT: The assessment is geared to individuals involved in analytics, including both business professionals and IT. If you are a consultant, please answer the questions with your most recent client in mind.

DEFINITION: For the purposes of this assessment, "analytics" includes traditional business intelligence as well as more advanced analytics such as predictive analytics, text analytics, and stream mining.

Analytics Maturity Model Assessment

Organization

This section focuses on ORGANIZATION related to your analytics efforts.

1. Preliminaries *

	Not even close/None	No there yet but moving on	Average	Active	Top 5%
How would you rate the current Analytics activity level in the company?	○	○	○	○	○
How would you rate the current Analytics competence level in the company?	○	○	○	○	○
How would you rate the current Analytics activity level compared to the markets/competitors?	○	○	○	○	○
How would you rate current Analytics competence level compared to the markets/competitors?	○	○	○	○	○
Expected company Analytics activity level in the next three years?	○	○	○	○	○
Expected company Analytics competence level in the next three years?	○	○	○	○	○
My current skill level/ability to use Analytics	○	○	○	○	○

2. Do you have both business and IT sponsorship across the company for analytics initiatives? *

- ○ We don't have sponsorship across the company for analytics
- ○ We have IT sponsorship only
- ○ We have business sponsorship only
- ○ We have both business and IT sponsorship
- ○ We have both business and IT sponsorship AND we work together
- ○ Don't know

3. Leadership *

	Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly Agree	Don't know
We are able to express the potential benefits of an analytics project in business language for executives to understand	○	○	○	○	○	○

4. Strategy *

	Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly Agree	Don't know
We have a well-established funding process in place for analytics. It is driven by both business and IT.	○	○	○	○	○	○
We have an analytics road map in place that has been agreed to across the company AND the discipline to change the road map if needed.	○	○	○	○	○	○
We take action using analytics (i.e., analytics as part of a business process, as part of a model) in my company.	○	○	○	○	○	○
Data and analytics drive our business in my company.	○	○	○	○	○	○

5. Skills *

	Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly Agree	Don't know
We train users to perform more advanced analytics.	○	○	○	○	○	○
We have users across the spectrum of analytic skill making use of analytics in my organization.	○	○	○	○	○	○

6. There are people in my organization with skills in advanced analytics to support the needs of the business (i.e., data scientists, business analysts, computer scientists, etc.). *

- No
- Yes, at the department or business unit level
- Yes, company-wide

7. Other things you want to point out related to Organization?

Infrastructure

This section focuses on INFRASTRUCTURE for analytics.

8. Analytics projects are driven by business leadership and deliver value incrementally instead of at the end of the entire development process. *

- No
- We are moving that way
- Yes, for some projects
- Definitely

9. Development *

	Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly Agree	Don't know
We have the right skills in place to address infrastructure technologies for our analytics efforts.	○	○	○	○	○	○
Our data scientists and analysts work closely with our data warehouse and data management teams to ensure that analytics workloads have the data infrastructure our employees need/use.	○	○	○	○	○	○

10. What infrastructure technologies do you currently utilize for your analytics efforts? *

- We use flat files or spreadsheets
- We have a data warehouse or a data mart
- We use an analytic platform or appliance
- We use a range of technologies, including our data warehouse, Hadoop, and others, but they are siloed
- We use a range of approaches that form an analytics ecosystem
- None

11. Do you make use of mobile technologies for analytics? *

- No, and we have no plans to do so
- No, but we are thinking about it
- Yes, but only for a select few
- Yes, for all those who need it

12. Architecture *

	Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly Agree	Don't know
We have a company-wide information architecture in place for analytics.	○	○	○	○	○	○
We have designed our architecture to take advantage of legacy systems already in place.	○	○	○	○	○	○

13. We use the public cloud for analytics. *

- ○ No, we would never use the public cloud
- ○ We have a private cloud that we use for analytics
- ○ We have a hybrid model where we use some public cloud, some data center, and some private cloud
- ○ We don't use the cloud yet for analytics, but we are moving in that direction

14. Other things you want to point out related to Infrastructure?

Data Management

This section focuses on DATA MANAGEMENT in support of analytics.

15. What kinds of data do you currently collect and manage as part of your analytics efforts? *

- ○ None
- ○ Structured data only from our internal systems
- ○ Structured data and demographic data
- ○ Multi-structured data along with our structured data
- ○ We collect and manage data from multiple sources, both internal and external to the company. This includes unstructured data, geospatial data, and much more

16. How much data are you analyzing currently? *

- ○ Megabytes
- ○ Terabytes
- ○ Petabytes
- ○ Don't know

17. We make use of multiple sources of data in a single analysis. *

- ○ No
- ○ Yes, with structured data
- ○ Yes, with structured data and 1 or 2 outside sources such as demographic data
- ○ Yes, with different kinds of data including unstructured data and other non-traditional data, but it is a hassle trying to integrate it
- ○ Yes, with different kinds of data and we do a good job of integrating it

18. Kinds of data *

	Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly Agree	Don't know
Employees at my company can easily find the data they need when they need it.	○	○	○	○	○	○

19. How do you integrate your data? *

- ○ We don't have a good way of integrating it yet
- ○ We have metadata that we use to help in data integration
- ○ We use a vendor's product for data integration such as data blending, unified information access, data virtualization in a data layer, or a logical data warehouse
- ○ We employ ETL routines to centralize as much data as possible in a data warehouse

20. Integration *

	Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly Agree	Don't know
Our data is stored in siloes.	○	○	○	○	○	○

21. If users want self-service access to shared data resources they can generally get it. *

- ○ No; users are limited to their own data silos and do not have access to shared data resources
- ○ Yes; we apply techniques such as data blending to enable self-service access to integrated data from multiple sources
- ○ Users, if they meet access criteria, have self-service access only to a centralized data warehouse
- ○ No self-service access, but through IT users can access some shared data resources.

22. Data Quality *

	Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly Agree	Don't know
We have a process in place for dealing with data quality that is dependent on the kind of data we are dealing with	○	○	○	○	○	○

23. Other things you want to point out related to Data Management?

Analytics

This section focuses on the status of ANALYTICS in your company.

24. What kinds of analytics techniques does your company use to analyze data? *

- ○ None yet
- ○ BI/OLAP tools, dashboards, reporting, and even real-time reporting
- ○ Those above as well as visual discovery
- ○ Those above as well as predictive analytics
- ○ Those above as well as other data mining or statistical techniques
- ○ We utilize all of the techniques described above as well as techniques such as social media analytics, geospatial analytics, text analytics, network analytics, or stream mining

25. Scope *

	Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly Agree	Don't know
Analytics are often automated as part of the business processes in my company.	○	○	○	○	○	○

26. We have a good idea of what business questions we are trying to solve with data in my company. *

- ○ Not yet
- ○ We are working on it
- ○ Yes, and we are trying to make it part of our culture
- ○ Yes, the questions are business-driven

27. Culture *

	Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly Agree	Don't know
We have tolerance for early failure with new analytics technologies in my company.	○	○	○	○	○	○
Analytics is seen as a competitive differentiator in my company.	○	○	○	○	○	○
We are able to deploy analytics to support performance management metrics so that users can more deeply analyze data associated with the metrics for which they are accountable.	○	○	○	○	○	○

28. How are analytics delivered in your company? *

- ○ We don't deliver them yet
- ○ The statistician or data scientist prepares and delivers them
- ○ IT or statisticians build a dashboard or other interactive delivery method
- ○ Business analysts or others create and distribute them
- ○ They are operationalized as part of a business process
- ○ We use a variety of distribution methods, including operationalizing and embedding analytics into a business process

29. What percentage of your organization's business analysts, data analysts, data scientists, and business users have the tools and know-how to analyze data in a self-service environment, without close IT involvement? *

- ○ 1-20%
- ○ 21-50%
- ○ 51-75%
- ○ 76-100%

30. Delivery methods *

	Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly Agree	Don't know
Business users who develop useful data visualizations or advanced analytics are able to work with IT to secure funding and focus resources on deploying analytics to other internal departments.	○	○	○	○	○	○

31. Other things you want to point out related to Analytics?

Governance

This section focuses on GOVERNANCE as it relates to data and analytics.

32. Data management and ownership policies are in place and documented in my company. *

- ○ No
- ○ We haven't had time for that but we know we need to do that
- ○ We are putting this in place now at the business unit level
- ○ Yes, at the business unit level
- ○ Yes, at the enterprise level

33. Structure, Compliance, Stewardship and Security *

	Strongly disagree	Disagree	Neither disagree nor agree	Agree	Strongly Agree	Don't know
We have an analytics governance team in place with representatives from across the company, including key business stakeholders. Roles and responsibilities are clearly defined.	○	○	○	○	○	○
We are monitoring adherence to our analytics policies.	○	○	○	○	○	○
The formalized role of the analytics steward is in place with roles and responsibilities are clearly identified.	○	○	○	○	○	○
Data definitions and metadata are clearly established.	○	○	○	○	○	○
Security policies are in place and enforced for all forms of data in my company.	○	○	○	○	○	○

34. Other things you want to point out related to Governance?

Demographics

This section focuses on survey DEMOGRAPHICS to help us benchmark scores.

35. Please select the answer that best describes your role at your company. *

- IT manager
- IT director
- IT executive
- Data Warehouse Manager
- Data Warehouse specialist
- Controller
- Analyst
- Business analyst
- Data scientist
- Business manager
- Business director
- Business executive
- Team leader
- Other

36. Sex? *

- Female
- Male

37. Age? *

- <20
- 20-29
- 30-39
- 40-49
- 50-59
- >60