LUT University

TUOTANTOTALOUDEN KOULUTUSOHJELMA

Data Warehousing as a Cornerstone for Successful Business Intelligence

Tietovarasto liiketoimintatiedon hyödyntämisen kulmakivenä

Bachelor's thesis

Johannes Jolkkonen

SUMMARY

Author: Johannes Jolkkonen

Thesis title: Data Warehousing as a Cornerstone for Successful Business Intelligence

Year: 2019 **Place:** Lappeenranta

Bachelor's thesis. LUT-University, Industrial Engineering & Management.
30 pages, 5 pictures and 0 attachments
Supervisor(s): Kirsi Kokkonen

Keywords: Business Intelligence, Business Analytics, Data warehouse, Database, Critical Success Factors, Project Management

Combining a literature review with action research, the objective of this thesis is to study the factors that play the biggest role in determining the success of a business intelligence implementation in a small enterprise. There is a particular emphasis on constructing the data warehouse as well as considerations in the ETL-process, with an aim to examine business requirements side by side with the technological matters that working with data necessarily entails.

What was found is a set of critical success factors that practitioners can apply to their BI-initiatives, as well as measures by which businesses can assess the degree of success in their BI-development projects. Also identified are the technological factors representing the most likely sources of difficulty, as based on the action research.

TIIVISTELMÄ

Tekijä: Johannes Jolkkonen

Työn nimi: Tietovarasto liiketoimintatiedon hyödyntämisen kulmakivenä

Vuosi: 2019 **Paikka: Lappeenranta**

Kandidaatintyö. LUT-yliopisto, Tuotantotalous.
30 sivua, 5 kuvaa ja ei liitteitä
Tarkastaja(t): Kirsi Kokkonen

Hakusanat: Business Intelligence, Business Analytics, Data warehouse, Database, Critical Success Factors, Project Management

Tämä tutkielma pyrkii kirjallisuuskatsausta ja toimintatutkimusta yhdistämällä selvittämään, millä tekijöillä on isoin rooli pienen yrityksen liiketoimintatiedon onnnistuneessa hyödyntämisessä. Tutkielma painottuu erityisesti tietovaraston suunnitteluun ja ETL-prosessiin liittyviin seikkoihin ja sen tavoitteena on tarkastella rinnakkain liiketoiminnan vaatimuksia sekä datan käsittelyyn vääjäämättä liittyviä teknologisia seikkoja.

Tutkielman tulos on joukko kriittisiä menestystekijöitä joita ammatinharjoittajat voivat soveltaa työskennellessään BI:n parissa sekä mittareita joiden avulla yritykset voivat arvioda omien BI-kehitysprojektiensa onnistumista. Lisäksi esitellään ne teknologiset tekijät, jotka toimintatutkimuksen pohjalta osoittautuivat kaikista haasteellisimmiksi.

CONTENTS

1 INTRODUCTION

Knowledge is power. This old adage is only becoming more true in our age of technology, as more data is being created than ever before and more tools are available for turning that data into information and knowledge. Every organization generates data as a part of its operations and they also have access to vast troves of external data, offering an invaluable window into performance, customer behaviour and market trends. Being as accessible and as invaluable as it is today, business intelligence and analytics have quickly become a central arena for companies to compete in. To benefit from this trend rather than getting swept aside by it, company management needs to open their eyes to the power that they have, much of it gathering dust within forgotten databases.

Many companies already hold more data than they know – data that could help executives make more informed decisions. What is needed is an effort to map out the data that is available, making it transparent to those who could benefit from it, and an investment in the technological tools that can make those benefits a reality. Yet, this process of transforming an organization to be more data-driven is rarely so straightforward in practice. Business intelligence initiatives often deliver disappointing results (Adamala & Cidrin, 2011), most often due to a separation between the business case that has been defined and the technical team responsible for its implementation (Yeoh & Popovič, 2016). This thesis will bridge that gap by showing how strategy and technology relate to one another when it comes to business intelligence.

1.1 Research objective and scope

This thesis sets to map out the biggest factors that determine the success of a business intelligence (BI) -implementation inside a small enterprise, where success is measured as the value created to the end-users of the system. There will be a particular emphasis on the Data Warehouse and the Extract, Transform, Load -process (ETL), and the underlying research problem can be divided roughly into two questions:

- How can companies measure and foster the success of their BI-initiatives?
- What technical considerations are involved in designing a Data Warehouse for BI?

This emphasis on data storage, transformation and flow has been chosen because these aspects are, from the researcher's experience, often insufficiently understood by people who deal more with higher level management and are less technically oriented in an organization, such as business analysts or executives. The researcher sees it as expedient to provide a digestible look into those technical aspects behind BI that ultimately enable the end-users to draw insights from data. The resulting thesis should hopefully allow business executives and other practitioners to start considering the potential value of BI in their organizations, dispelled of common misconceptions and aware of typical blunders in its implementation.

1.2 Research methodology

This study consists of roughly equal parts of literature review and empirical research. The primary driver behind this research is the researcher's employment to a BI-project in a company that wants to modernize how they utilize data. The first part of the project revolves around defining the desired outcomes and planning the necessary technical architecture, which will be done through semi-structured interviews with relevant stakeholders. This part will also serve as action research for the purposes of this thesis.

Action research is a variant of the case study method, characterized by the researcher's role as an active participant in solving the researched issue rather than as a neutral observer (Baskerville & Wood-Harper, 1996). It is an appropriate method for investigating problems or processes that are cyclical and iterative in their nature (Coughlan & Coghlan, 2002), and thus a good fit for information systems (IS) research. The action research will be guided and supported by a literature review, during which I will build a general understanding of BI, exploring frameworks and best practices for leveraging it for most value possible.

1.3 Thesis structure

The first half of the thesis will be theoretical, consisting of three sections: a definition and overview of what BI is to begin with, followed by an explanation of the different components in a typical BI-architecture and finally, a framework for critical success factors and success metrics that can be applied in the context of real BI-initiatives.

The second half of the thesis will be empirical, introducing the company where the action research was carried out and then working to define the business case behind their BI-initiative. After defining the business case, I will propose an appropriate solution for data warehousing and data loading. I will then finish by answering the two research questions based on both the literature review and the action research, along with sharing any other afterthoughts.

2 THEORETICAL BACKGROUND

The theoretical section of the thesis consists of three parts: a definition and overview of what BI is, an explanation of the typical BI-architecture and finally, a framework for critical success factors and success metrics that can be applied to BI-initiatives.

2.1 Overview of Business Intelligence

Business Intelligence (BI) is an umbrella term that spans people, processes and tools to organize information, enable access to it and analyse it to improve decisions and manage performance (Chandler et al., 2011). To put that in simpler terms, BI is about using technological tools to create knowledge out of data and making that knowledge accessible in order to make better business decisions and ultimately drive an enterprise's success.

The annual study by Gartner to track the top priorities of company CIOs has consistently seen BI and analytics take home the number one spot since 2012 – typically by a landslide margin – and only being abruptly displaced by artificial intelligence in 2019 (Gartner, 2013; 2019). As for market size, BI was globally valued at $17.15 billion in 2016 with a projection to reach $147.19 billion by 2025, giving a compound annual growth rate (CAGR) of 27% between 2017 and 2025 (Marketwatch, 2019).

This indomitable presence and the figures of booming growth in BI over the past decade are marked by an understanding that data is now too readily available and its value in decision making is too great for it to be ignored by any modern enterprise that wants to remain competitive. Beyond just growth, there's been a recent but significant trend within the industry towards self-service, enabled by the proliferation of access to cloud platforms (Gartner, 2017). This signals a move towards making powerful data analytics available with an increasingly lower threshold, both in terms of financial resources and technical know-how.

2.2 Layers of a Business Intelligence -architecture

Even with rising trends, the models that are used in the literature for a typical BI-architecture remain largely similar to the five-layered framework proposed by Ong et al. (2011), as pictured

in Figure 1 below. These five layers are the Data Source, ETL, Data Warehouse, End User and Metadata layers, and in this section I will explore each of them in more detail.

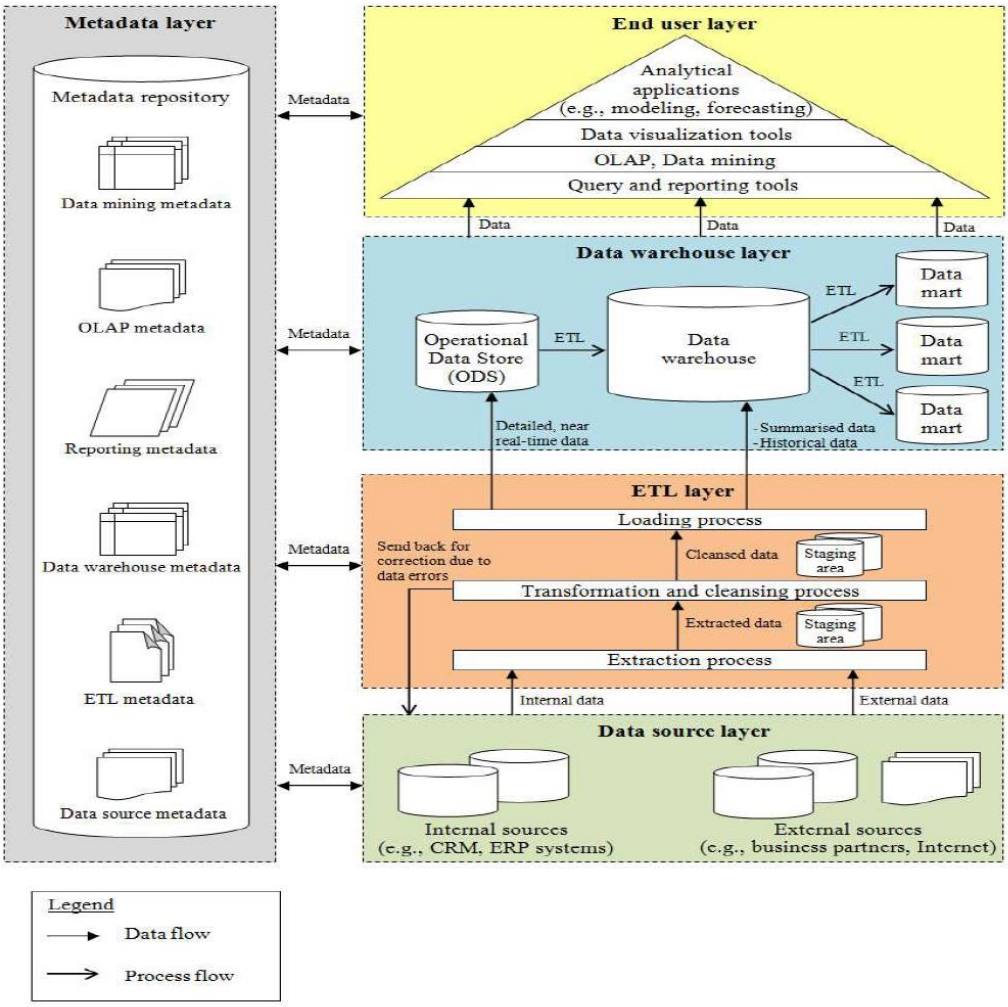

Figure 1: General model for BI-architecture (Ong et al. 2011)

2.2.1 Data Source -layer

Companies can acquire their data from either internal or external sources. Internal sources refers to data captured and maintained by operational systems inside an organization, such as Customer Relationship Management- (CRM) and Enterprise Resource Planner- (ERP) systems. Data from internal sources thus relates directly to both front-office and back-office business operations, e.g. customer interactions, sales, transactions and accounting information (Klie, 2015). External data sources refers to data that originates outside an organization. This type of data can be collected from sources like business partners, the Internet, and public institutions.

This is data that usually relates to the conditions of the market and the competitive environment (Ong et al., 2011).

A comprehensive mapping of data sources is important for organizations because it allows data to reliably be obtained to meet business requirements, whether recurring or emergent. In addition to knowing where specific data can be obtained, it's essential to also understand how data is formatted across different sources in order to avoid problems in data warehousing and analysis (Davenport & Harris, 2007).

2.2.2 ETL-layer

The three processes of extraction, transformation and loading are the three components that make up the second layer. Firstly, extraction refers to the selection of data that is relevant for decision-making from among all the data that is available in the identified data sources, both internal and external. The extracted data is temporarily stored in what is called a 'staging area', after which it is transformed and cleansed as necessary (Turban et al., 2013).

Transformation refers to the conversion of data, by methods such as aggregation, to formats that are useful and consistent for analysis, while cleansing refers to the correction of errors in the data (Bose, 2009). Thirdly and lastly, the data is loaded into a target repository, such as a data warehouse. The staging area has an important role throughout the ETL-process, because temporarily storing the data after each intermediate phase can prevent the need to do repeat work in case the process unexpectedly fails or terminates (Ong et al., 2011).

2.2.3 Data Warehouse -layer

The Data Warehouse -layer introduces us to three concepts of data storage: the operational data store (ODS), the data warehouse (DW) and the data mart (DM) – each of them distinct in how they store data. The ODS can be thought of as an intermediate, volatile and frequently updated storage that's used for bringing together all the data from different sources through the ETL-layer. From the ODS data can then be passed onwards, either directly to reporting applications or to another, more long-term storage repository (Turban et al., 2013; Ong et al., 2011).

DW, The layer's second component, is used exactly for this type of long-term storage. One of the most important parts of a BI-architecture, the DW stores aggregated or summarized data, including historical data for long term analysis. The DW is characterized by non-volatility, meaning that while new data is added regularly, existing data doesn't get overwritten or deleted. Thus, where data in the ODS are usually stored for a maximum of 60-90 days, for DWs it can be 5-10 years (Chan, 2005). Finally, DMs are subsets of a DW that are used to support the analytical needs of a particular business function, whereas the DW is enterprise-wide (Ong et al., 2011).

2.2.4 Metadata -layer

Metadata refers to data about data, such as its storage location, source and the changes made to it – all of it usually aggregated into a single metadata repository. In addition to technical information, metadata stores also contain business rules and definitions that have been defined for the data, informing users of its different manners of use (Davenport & Harris, 2007; Turban et al., 2013). In other words, good management of metadata can not only reduce the time spent on development and maintenance drastically, it can also make the system much more user-friendly through the data-related information that it provides for its end-users.

2.2.5 End User -layer

The top layer of the architecture consists of front-end tools that show data in different formats to different users. These tools, pictured in a pyramid shape, range from ad-hoc queries to advanced visualization and analytics tools with functionalities such as real-time dashboards, forecasting and scenario modelling (Gartner, 2017; Turban et al., 2013). As one moves up the pyramid, data gets presented more comprehensively to serve the growing complexity in decision-making, reflecting the management hierarchy among users of the different tools (Ong et al., 2011).

To close this overview of the BI architecture, I will note that as the organizational structures of different businesses vary, so will the ways in which they should implement the layers discussed here. To give you an example; for a business whose source systems provide data in a format that is immediately useful for their analytics, their respective ETL process might have a very

minimal or even non-existent transformation function. Another business where different business functions have largely the same data needs, data marts could well be omitted while relying on just a single, shared data warehouse – particularly if the business is relatively small.

2.3 Critical success factors and success criteria for BI

Most BI and data warehouse projects fail, with a failure rate that has been estimated to be as high as 80% – this was the conclusion of a literature review by Adamala & Cidrin (2011). Even large investments in BI initiatives have historically yielded little to no benefits for the organizations implementing them (Williams & Williams, 2007), indicating that the challenge is not one that can be fixed simply by more dedicated resourcing.

Part of the problem is that for most of its lifetime, BI has suffered from a void of generally accepted critical success factors (CSF), both in practice as well as in the academic literature (Yeoh & Popovič, 2016) – yet just as it is with any initiative, lacking clear measures for success makes it less than likely that success will be achieved. Furthermore, the benefits that BI has to offer for decision-making are often subjective and hard to measure in monetary terms (Sangar & Iahad, 2013). Then, even if benefits are successfully realized, it can be challenging to weigh them against the investments that have been made.

To address these challenges, Yeoh & Koronios (2010) came forward with a three-dimensional set of CSFs for implementing BI systems (Figure 2), derived from studying real-world organizations and further validated in another case study by Yeoh & Popovič (2016). This framework, which I will use as my main guide through the rest of this thesis, provides insight into the characteristics of BI in practice. Consequently, it then offers a set of success criteria that practitioners should use when embarking on their own BI-projects.

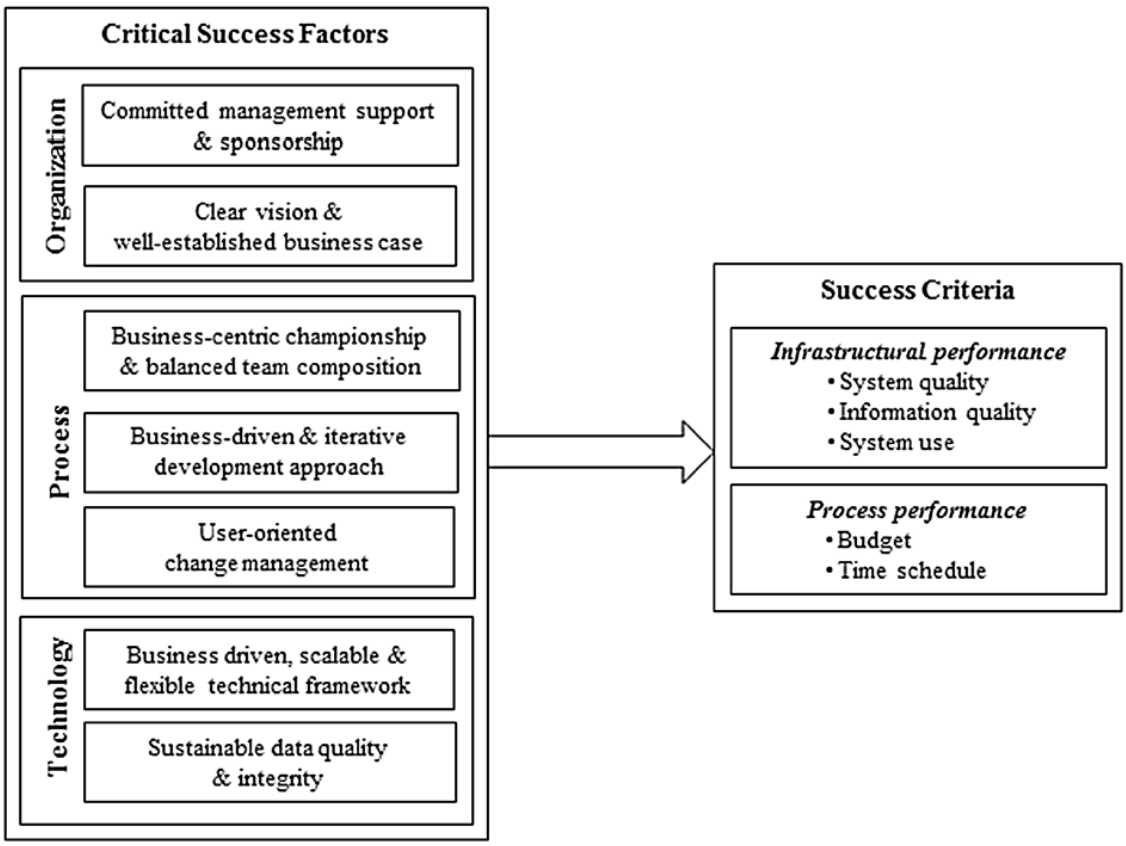

Figure 2: Critical Success Factors for BI-implementations (Yeoh & Koronios 2010)

2.3.1 Organizational dimension

Dedicated support and sponsorship from management has been widely recognized as the single most important factor for a successful BI-implementation (Yeoh & Koronios 2010). Sponsorship in the context of project management refers to the sufficient provision of resources and support for the project, as well as the delegation of accountability for the project's success (Project Management Institute, 2017). The ideal sponsor for a BI-project should come from a business function, because they are then more likely to have a strong stake in the success of the project (Watson et al., 2001).

As a BI-initiative is driven by business, it is essential from the onset to have a clear vision for the initiative, comprising a well-defined business case, its requirements and its objectives (Yeoh & Koronios 2010). This business case should be the north-star that moves the implementation forward, for otherwise the end result is unlikely to be embraced and adopted by the end users.

Information system -related projects are especially known to face strong organizational resistance to change (Kim & Kankanhalli, 2009), and so a “build it and they will come” -style approach is one that should be avoided.

2.3.2 Process dimension

Moving down from the level of the organization, business-centricity should also be seen at the level of the process. Echoing the case for selecting a project sponsor from a business function, the project should similarly have a champion with business acumen. This champion role is more connected to the project’s hands-on -execution than the sponsor, but still not as closely as an actual project manager (Project Management Institute, 2017). Instead, the champion’s primary responsibility is to continue steering the course of the project as organizational challenges arise, keeping the end-users’ needs closely in mind all the while doing so.

The process dimension also highlights the iterative nature of BI (Yeoh & Koronios 2010). There are a couple of reasons for this. Firstly, managing a large-scale organizational change always entails serious risks, given the amount of variables that need to be considered concurrently (Ang & Thompson, 2000). Secondly, modern businesses are subject to rapid change in any case, both with regards to their environment as well as to their internal decision processes (Cook, 2015). It is also often the case with BI-systems in particular that the needs of their users end up shifting and evolving after they get their hands on the system and learn more of its capabilities. Consequently, it can be difficult to accurately predict in advance what costs and workload is needed for development and maintenance in the future (Fuchs, 2004). Implementing BI is not then a project so much as a process, and the general consensus is that it is better approached with incremental improvements rather than attempting to roll-out a finished solution right off the bat.

Following this reasoning, an incremental development approach both mitigates the risks above and gives the stakeholders a more steady demonstration of progress (Dobrev & Hart, 2015). A change that happens gradually is easier to manage than a sudden one, and small initial benefits become evident and act as a catalyst for accelerating further development. Those first implementation steps should thus be chosen carefully, ideally within a high-impact business area with generous room for process improvement (Bose, 2009).

Having talked at length about the importance for BI initiatives to be business-driven both at the organizational and the operational level, it seems wise to insert here a quick reminder of the technical side that working with data necessarily entails. Even if the underlying business case is clear and valid, it is more than likely to degrade on the way to its fulfilment if not for a cross-functional team that can see it safely through. The team needs to have sufficient technical expertise in areas such as data engineering and software testing in order to build a system with the desired functionalities. In the section that follows, I will cover the framework's technological dimension in more detail.

2.3.3 Technology dimension

The selection of technical tools should accommodate the tendency of BI-systems to grow larger in scale than can easily be estimated beforehand. To ensure the system's long-term viability amidst dynamic and expanding business needs, software and hardware choices should favour those that offer great compatibility with legacy systems and flexible options to scale hardware resources like storage space and processing power (Olszak & Ziemba, 2007). While those might not be as straightforward or cheap to initially implement as some turn-key solutions, the risk of having to reconsider the entire technological architecture mid-development seems worth mitigating.

The other main technical consideration concerns data integrity and quality, as previously hinted at under section 2.2 when discussing the data source and ETL -layers of the BI architecture. If inaccurate data are initially brought to the system or errors arise due to the way data is formatted, there is no possibility of drawing reliable insights from the data further down the line (Yeoh & Koronios, 2010). This sentiment is quite accurately summarized in the phrase "garbage in, garbage out", popular especially among software engineers.

Aside from carefully mapping data sources to identify and correct corrupt data, it is also essential to have common measures and definitions across the organization (Turban et al., 2013). While this might seem obvious, it is not unusual for different business units to define terms in ways that best suits their own purposes, resulting in terms having slightly different meanings to different people. So, as BI-systems are typically cross-departmental, this type of

ambiguity can easily impair communication between departments and even lead to misunderstood reporting and erroneous decisions. A good way of preventing these terminology problems from the onset is to establish an enterprise-wide set of definitions that the BI team can use when constructing the data warehouse (Yeoh & Koronios, 2010).

Many issues in the back-end and source systems aren't discovered until data are populated into front-end BI systems, at which point it will potentially influence decision outcomes (Watson et al., 2004). Not only is it less than ideal to discover issues by making business errors, correcting these issues also gets increasingly more complex as we move downstream from source systems to the front-end analytics (ibid.). Putting in the effort from the ground-up to ensure that source data is well understood is generally well worth it, seeing how it can subsequently prevent the need to correct data flows and transformations that are already running across the entire architecture.

2.3.4 Success criteria

In the context of these three dimensions of CSFs for BI, Yeoh & Koronios (2010; 2016) give their metrics for measuring performance in two aspects – process performance and infrastructure performance. The first aspect refers to the process of implementation, while the second one, infrastructure, considers the resultant system and its quality.

Process performance can be measured in terms of schedule and budget considerations. Schedule brings into question the time that it takes to implement an initial version of the system, whereas the budget criteria bring oversight into the cost of developing and maintaining the system (Ariyachandra & Watson, 2010). What exactly constitute appropriate schedules and costs for a given project can obviously only be defined per-case, but what matters is that they are defined before the onset of the project.

Infrastructure performance here is based on three dimensions of the information system success model by DeLone & McLean (1992, 2002), namely, system quality, information quality and system use. System quality reflects the system's performance metrics, which can be seen in measures like availability, reliability, and response times. Information quality refers to the accuracy, completeness, timeliness, relevance, consistency, and usefulness of information

provided by the system. Together, system and information quality are the most common dimensions along which information systems are evaluated, because they directly impact both user satisfaction and the intentions that the user develops for using the system. (Popovicˇ et al., 2012). Lastly, system use is concerned with the manner in which users consume the outputs of a system. In the context of BI, these outputs are the various forms in which data is presented to users, and consumption of these forms enable them to draw business insights from the data.

Sustained and increasing system use can be considered a good metric of IS-success in the long term, as it reflects the value that brings people to continue using it (DeLone & Mclean, 1992). In this, ensuring the underlying system and information quality plays a big part. Kim & Kankanhalli (2009) also highlight the training of users as one of the most effective ways to increase user activity.

2.3.5 Conclusion on BI success factors

The primary characteristic that separates BI from other common information systems is that while operational systems such as the ERP and CRM are centered around well-established business-processes, the use of BI evolves continuously over time. This is because reporting and analysis are rarely pre-defined processes and business users are usually free to conduct them as they like. This characteristic ripples over the whole BI-initiative, giving rise to the philosophy of iterative development that should be present at all levels, from project team composition to technology decisions.

It also bears repeating that among all the CSFs discussed above, the organizational dimension was found to be more significant than any of the rest. The takeaway then follows; before any other steps are taken, company management should first sit down and specify the exact needs they are looking for BI to solve. Planning the nature and scope of the implementation effort can then be done based on those needs, and it should be done in such a way that the implementation comes to enjoy committed support from management. A concrete schedule should also be defined before starting to build the system. In the spirit of incrementalism, milestones under the schedule should be modest and specific, rather than overly ambitious or broad.

The BI-team, composed of both technical and less technical members of the organization, should be in continuous contact with end-users. Only through this contact can the team ensure that the system develops into the proper direction and provides sufficiently user-friendly service – this includes technological quality metrics like reliability and response times, but also the quality of information; relevance, accuracy and completeness. As the system becomes more and more usable, encouraging its users also becomes increasingly important – not only continuing to listen to their feedback but also training them in the system's use.

3 CASE: BI-IMPLEMENTATION AT CASECO

With this section I will go into the empirical part of the thesis, first introducing the company where it was done and then describing their business case for BI. For this business case, I will map out the company's main sources of data and define the analytical needs for the user group.

3.1 Overview of the company and the BI-project

The case company, thereafter referred to as CaseCo, is a small company with fewer than 40 employees. CaseCo operates in the finance industry, serving a global B2B-clientele by offering invoice factoring as their primary service. Invoice factoring is a way for businesses to accelerate their cash cycle by selling their invoices to a third party at a discount, receiving an immediate advance on their receivables in return. As well as being able to outsource their credit risk to the third party, businesses also benefit from invoice factoring by freeing more of their tied-up capital for investments in their business objectives.

With CaseCo's management group as end-users, the primary objective of this BI-project is to modernize the use of data in decision-making and reporting. In practice, this means driving up the degree of automation for tasks that are currently done manually, such as searching, loading and visualizing data. Currently, a considerable amount of data needs to be queried from an enterprise database by IT on behalf of management, which creates serious bottlenecks in reporting. What is hoped for is a system that provides each member of management unobstructed access to the data they need, allowing business insights to be drawn more effortlessly and, where relevant, in real-time. The secondary objective to be carried out concurrently is to deconstruct the metrics and business definitions that are currently used by management. In doing so, we will clear any inconsistencies as well as finding ways in which those metrics could be improved or appended to.

While the goals for the BI-implementation are described here in their entirety to give context, this thesis will only report on that portion of the project that concerns defining the business case and the subsequent construction of the data warehouse and its ETL-flows.

3.2 Overview of available data sources

CaseCo's primary data source is a database which in turn is populated by data from the two enterprise systems that the company uses – I will call them System A and System B. There are also other, more fragmented sources of useful data, most of them related to marketing.

3.2.1 Enterprise systems and database

System A is used for the processing of incoming invoices, and as most of CaseCo's activities and revenue revolve around these invoices, the system essentially plays the role of an ERP. Receiving, processing and either purchasing or rejecting invoices is carried out on the system by CaseCo's customer service department and as a result, the system stores comprehensive data about each individual invoice; data such as its registration date into the system, due date as well as the payment's status, value and associated share of commission. When changes are made to the status of an invoice, the system also tags those changes with the employee that made them.

System B is a portal used by CaseCo's customers and sales personnel, and it can be likened to a CRM. While customers can use it to manage the status of their invoices, the sales team uses it to see basic information about each customer as well as CaseCo's history of correspondence with them. System B is used to enter and update data about customers, such as their contact details, contract type and selling activity. The process of customer acquisition is also tracked on System B with details about sales calls, meetings and contracts made.

Tapping into these two systems, the enterprise database then contains all the core operational data of CaseCo, including records of its bank transactions.

3.2.2 Miscellaneous data sources

There are also a handful of data sources that are unrelated to the database discussed above. The most important one of these is a dialling software used by telemarketers and customer service, generating data on the calling activity of each user. Social media platforms used by CaseCo for marketing provide their own data about traffic and user engagement, and Google Analytics does the same for the company website. CaseCo also uses a Voice of Customer -service (VOC) for surveying customer experience and satisfaction, which offers complete analytics dashboards on

its own. Finally, they have a 3rd party CRM that is currently used for a handful of the company's largest customers. These secondary sources vary greatly in their degree of relevance and all of them are accessed in their own separate environments.

3.3 Defining analytical needs for the end-user group

CaseCo does most of its internal reporting on a monthly basis. Reporting is a combined effort from the five members of its management group, namely the company's CEO, Financial Controller, Risk Manager, Customer Service Manager and Marketing Manager. Each of them focuses on a specific group of key performance indicators or KPIs, although there is some overlap within their reporting. Interviews were carried out with each of them to determine what metrics are the most vital and what parts of their reporting involve the most manual work.

CEO

The CEO of CaseCo is primarily concerned with the value and maturity of the company's portfolio, that is to say, the outstanding value of purchased invoices and the average time in which their payments are received. Changes are tracked monthly to see trends and seasonality in the portfolio's value, correlating them with annual revenue forecasts and budgets. Because revenue is ultimately driven by customer acquisition, the CEO also keeps a close eye on lead generation as well as contracts that have been won or lost, emphasizing on the company's most high-profile customers.

Financial Controller

The controller manages all of CaseCo's transactions, from accounts payable and payroll receivables to general operational finance. This also means keeping track of changes in the company's cost structure, capital structure and profitability, as well as drawing forecasts for profits and liquidity. Transactions with international customers were found to be especially laborious for the controller, as VAT needs to be manually calculated for each of them. This issue is aggravated by the fact that presently it is not possible to find all of those international transactions at any one place.

Risk Manager

In accordance with managing overall risk and minimizing the potential of credit losses to the company, the risk manager tracks that share of CaseCo's receivables which is overdue and determines individual risk ratings for each of the company's customers. This risk rating is determined by factors such as their credit rating, the reliability of payments from that customer's invoices in the past and sudden changes in their activity or volume of invoices sold. These metrics are automatically calculated with Excel and involve relatively little manual work. However, risk ratings in particular have to be looked at individually to track changes and identify critical customers. It could be beneficial to have a way to make them more readily apparent from the data, as opposed to the simple tables that are used to do this identification job now.

Customer Service Manager (CSM)

The CSM manages CaseCo's customer service department, responsible for the reception and purchase of invoices as well as debt collection. The CSM reports on the volume of invoices received and purchased, as well as that portion of them which is rejected by CaseCo due to perceived risks of credit loss or fraud. All of these volumes are tracked on the yearly, monthly, weekly and daily intervals, allowing for the tracking of the departments efficiency in terms of processed invoices per work hour.

Because the department also deals with debt collection and payment reminders, it would be helpful to have a view of all overdue invoices in a single file, grouped by customer. As it is, they can only be viewed individually or by their total sum. CaseCo would also like to start tracking the department's efficiency by employee. Data on the activities of individual employees is stored by System A as mentioned in section 3.2.1, but it is currently not accessed.

Marketing Manager

Marketing tracks CaseCo's customer acquisition per channel and the effectivity of its advertisement campaigns. This means comparing the amount of inbound leads the company gets from different sources, as well as their rate of converting into customers. The company's main sources of leads by order of importance are outsourced telemarketing, the company

website and social media. While the effectiveness of telemarketing providers is currently tracked, the company aspires to do it in a more continuous and visual manner, as opposed to the periodical, ad hoc way it is done now. Furthermore, as all of the different metrics used for marketing are viewed quite separately from each other, the marketing manager would like to start grouping them in a more meaningful way, such as by brand strength, inbound and outbound.

3.4 Conclusions on the business case

Interviewing members of management, it was found that CaseCo's operational diagnostics are driven above all by metrics around customer acquisition. Secondly, it is important for the company to have a hold on the activity levels of its customers; while a customer might have an active contract, they can still freely choose how many of their invoices they sell. This produces a challenge in predicting revenues, and so there is great value in being able to understand and influence the customers' selling behaviour. Thirdly, the company wants to know how efficient their sales and customer service staff are in contributing to these first two areas.

While Excel templates are used to some degree by all users for deriving KPIs and creating report tables and graphs, they all also share the challenge of having to go to IT for at least some of the data that they want to enter into those templates. In other words, easier access to the same data that is already being used would already bring a lot of value all by itself. While that will be our first task, we will also add to that familiar data any new pieces that can give more insight into the three performance areas identified above.

4 CONSTRUCTING THE DATA WAREHOUSE AT CASECO

The enterprise database described under section 3.2 is hosted at the remote location of CaseCo's parent company. Due to the sensitive nature of the data, access controls to the database are strict and currently only CaseCo's IT department can access it. To alleviate the bottlenecks this creates in data availability and to enable the effective implementation of BI down the line, our first step is to construct a data warehouse that's hosted at CaseCo's own offices and is then more accessible for applications down the line. In this section I will go through this process, from the choice of technologies and data sources to the design of ETL-pipelines.

4.1 Technology decisions

After exploring a variety of alternatives, Microsoft's SQL Server (MSSQLS) was chosen as the software for building the DW. The primary reason for deciding to build the DW as on-premise rather than using a cloud solution was security; because data pipelines will be automated to independently load data from the parent company -hosted database, a cloud solution would likely have to be accompanied by particularly rigorous reassurances and access controls. Another reason behind the choice of MSSQL is that the software license includes SQL Server Integration Services (SSIS). As an extension to Microsoft's Visual Studio, SSIS is a powerful set of tools for building and automating ETL pipelines on a schedule. MSSQL also stores metadata (see section 2.2.4) for the server automatically and allows for the easy management of user permissions.

4.2 Prioritization of Data Sources

The enterprise database contains data not only from CaseCo's operations, but also the operations of other branches under the parent company – most of it irrelevant for the purposes of CaseCo. Keeping in mind the iterative nature of BI development as described under section 2.3.2, we will start by bringing in only the most foundational operational data, then adding to it gradually. As well as mitigating project risk, this also works to ensure that all the data imported to our new DW remains well-documented and useful. By starting with simple and basic operational data and metrics, we can get users to engage with the new BI-system quicker and make its successful adoption more likely. There are four tables in the database that contain most of the data used by management; one for data about customers, another for invoices, third for

bank transactions and a fourth for employees – these are what will initially be imported into the DW, and I will touch more on data tables in the section that follows this one.

Meanwhile, some of the secondary data sources listed under section 3.2.2 can be used to track more closely the generation of new leads, which was identified as a high priority activity – telemarketing in particular generates the lion's share of CaseCo's new prospects and customers. Among the secondary sources we will then prioritize data generated by the dialling software, importing it to the DW as soon as possible so it is available for analysis down the road. It might also prove useful to import external data about social media reach and website traffic into the DW, as to reduce the number of different sources that marketing analytics has to use. While the 3rd party CRM might be expanded to manage a larger proportion of CaseCo's customer's in the future, importing data from it at this time was not seen as an immediate priority.

4.3 Relational tables and data groups

Figure 3 below portrays a simplified database diagram that represents the relevant portion of our source database. In the diagram, you can see four tables – Customers, Invoices, TransactionEvents and Employees –, the columns of data they contain and the data type of each column. Knowing these data types matters a great deal, because trying to combine data of different types is a common cause for the break-down of ETL-pipelines.

The arrows in the diagram indicate relations between the tables, coordinated by matching Identifiers or IDs. Each row in the Invoices-table, for example, contains a customer ID which connects it to a certain customer in the Customers-table, as indicated by the green arrow. Likewise, the row contains under its column 'registered_by' the ID of that employee who registered the invoice into the system, and this connection is indicated by the orange arrow. Databases that contain data in this format of interrelated tables are called relational databases.

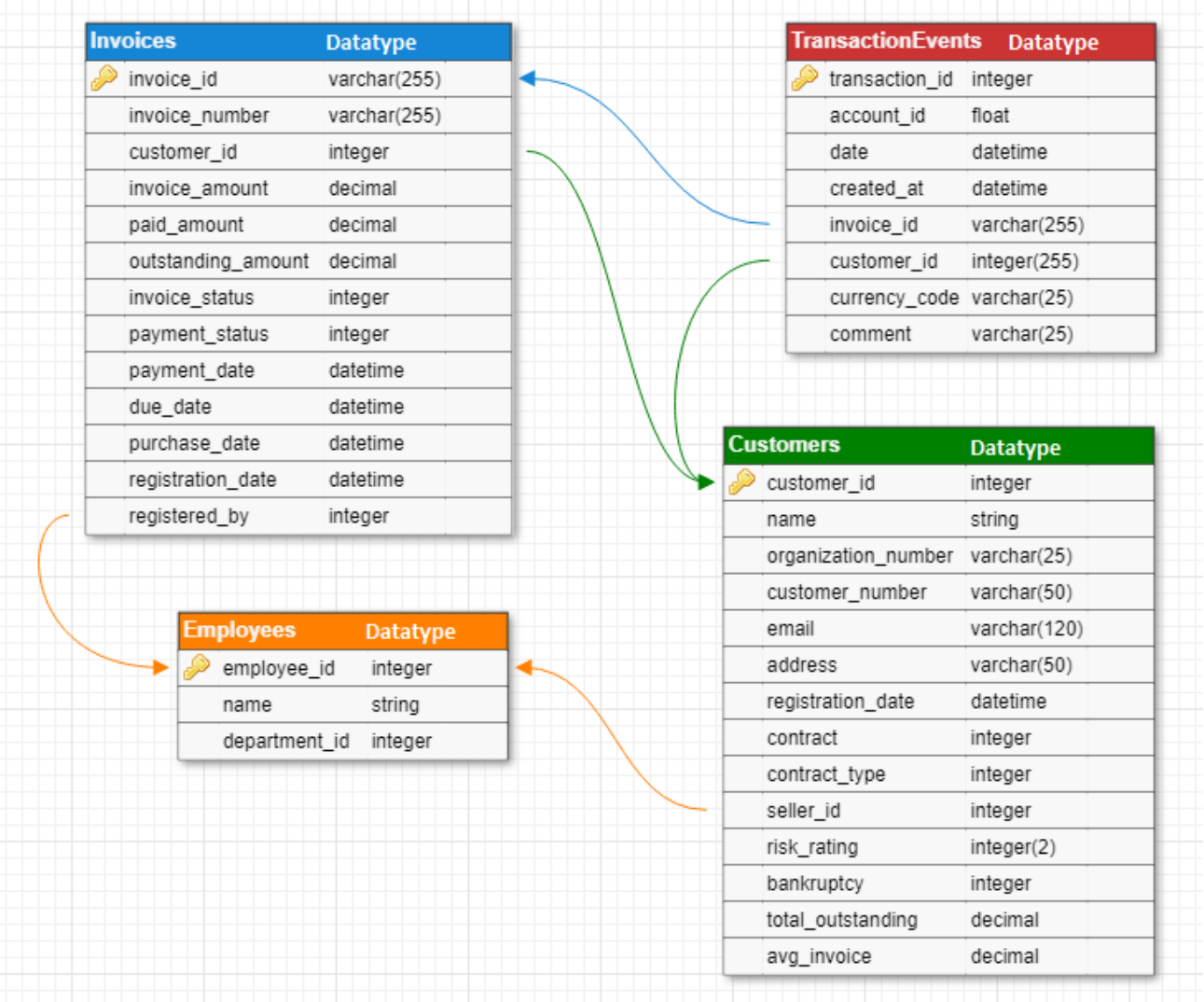

Figure 3: Database diagram with four relational tables.

The tables in the actual enterprise database have many more columns than are included in this diagram, and more than would fit into any intelligible picture for that matter. However, not nearly all of the columns are useful for analytics. This diagram illustrates that core selection of them which will first be loaded to our DW.

4.4 Designing the ETL-processes

Now that we have established what data to load into the DW and how that data is formatted, I will explain the basic ETL-process that will be implemented using SSIS. What is shown here is a process called an incremental update from a source table to a destination, where SSIS looks

for new rows or changes to rows within the source and then updates the destination accordingly. In Figure 4 below is shown a high level flow diagram of this process.

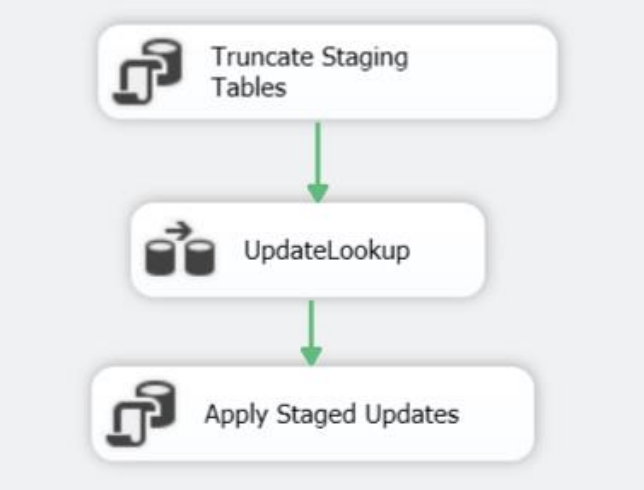

Figure 4: The control flow of an incremental update task in SSIS

To make more sense of the logic behind the process, Figure 5 shows you more closely what happens within the data flow task named UpdateLookup.

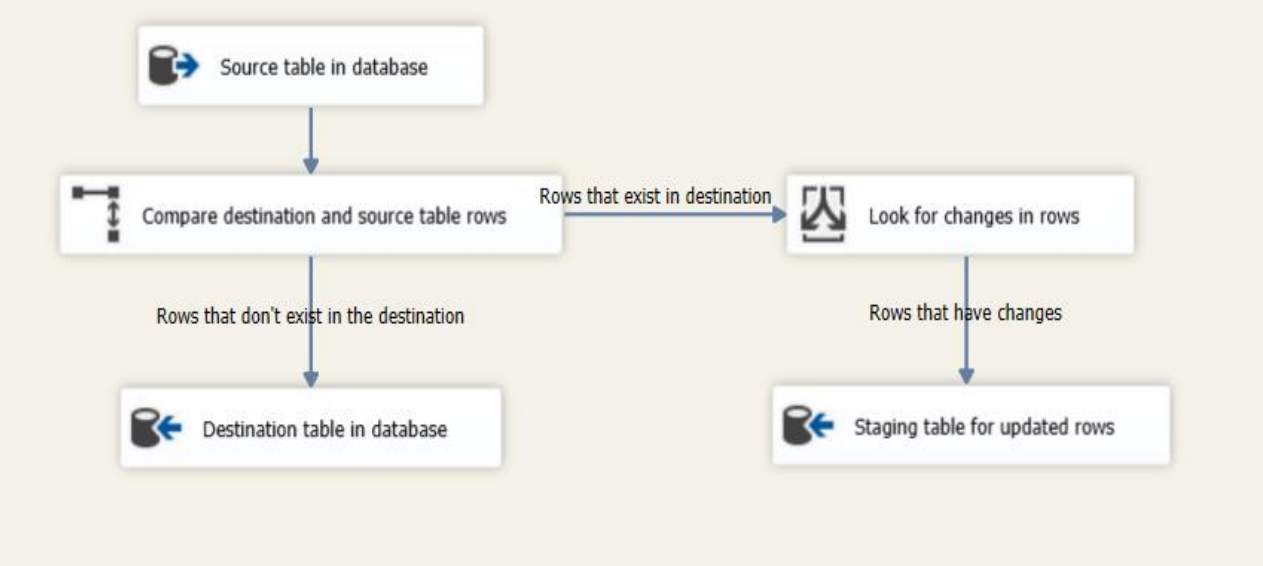

Figure 5: The control flow of an incremental update task in SSIS

Starting from the left side of the figure you can see the source and the destination tables, connected by what is called a lookup task. By comparing row IDs, the lookup finds rows in the source that do not yet exist in the destination and subsequently loads any new rows it finds into the destination. Meanwhile, any rows that match the ID of a pre-existing row in the destination are sent to another task to the right. What this second task does is compare the matching rows

for any differences apart from their IDs, and stores the differing rows from the source table into a staging table in the bottom right corner.
Back to Figure 4 and assuming the source table contains the most recent information, what follows UpdateLookup in the flow diagram is a script task. This script task, written in SQL, replaces any outdated rows in the destination with their updated counterparts from the staging table.

As a side note, a staging table is used here for two reasons. First one has to do with performance, because using SQL circumvents the need for SSIS to compare and update the rows one by one, which can take quite long. A staging area, as you might recall from section 2.2.2, also provides an additional place to check for clues in case something unexpected happens during ETL execution. The staging table is, however, emptied out by another SQL script at the beginning of each execution so that the program can be run repeatedly.

Constructing a process like this one for all four tables in Figure 3, we can keep our DW up to date as data continues to change in the operational source database. SSIS allows for these processes to be deployed as 'packages' and executed on a schedule, so the data can be set to update as often as necessary. As the data in our four tables is used for daily and monthly reporting rather than needing to be tracked continuously throughout the day, it will be set to update nightly.

While this simple process will give us sufficient data to start testing the system's reliability and get users to start experimenting with BI-applications, it hardly demonstrates the T in ETL – yet. Data transformations will undoubtedly be included in future SSIS jobs to correct inconsistencies in data, improve information quality and possibly to derive business KPIs directly from source data.

4.5 Roadmap for BI-development

Before closing the scope of this thesis, I will add some final thoughts on the future development of CaseCo's BI project. With an initial version of the DW in place, the next thing to do is choose a suitable front-end BI-application for working that data into actionable insights. After the application is in place, an investment in training users in how to use it would probably reduce the time it takes them to start engaging with it more. As per the CSFs discussed in section 2.3,

the importance of user engagement is obvious not only as the ultimate driver of value for the company and its management, but also as a way to get enough feedback throughout the development process.

Building on top of those few foundational data pipelines in the previous section, a roadmap will be set for each consecutive addition of data. This roadmap must come with a pre-determined schedule and at least a rough budget, as to be able to evaluate the implementation effort in line with the success criteria discussed in 2.3.4. Aside from continuing to fill in data to the four basic tables as necessary, data for telemarketing activity will be treated with high priority on this roadmap. That being said, communication with the end-users remains paramount, and if some new, clear requirements should emerge from among them, roadmap priorities should be able to accommodate for it.

5 FINDINGS & CONCLUSIONS

In this final chapter I will conclude an answer to the research questions, reflecting on the interplay between the theoretical and the empirical parts of the thesis. I will also talk briefly about the limitations that this thesis and its results are subjected to.

5.1 Answering the research questions

The research problem of this thesis was split into two questions:

- How can companies measure and foster the success of their BI-initiatives?
- What technical considerations are involved in designing a Data Warehouse for BI?

For the first question, an exploration of critical success factors provided strong grounds for the importance of organizational commitment and the continual involvement of business stakeholders in BI-development. This motivated an approach to CaseCo's project in which end-users were interviewed first before making any technological considerations, and this way it was confirmed that the management group indeed has a very strong and specific business case for a BI-initiative. Judging by the extent to which subsequent technical planning was made easier by this, it seems to have been a good approach.

We will assume that a continued emphasis on user engagement, information quality and the other measures of infrastructure performance will foster the initiative's success in the long run, too. Emphasis on these things can be achieved in part by taking an iterative approach to BI-development, which was validated by the theory and influenced decision-making in the empirical business case. Namely, loading operational data to the data warehouse will be done incrementally, starting from a selected group of foundational data points and then gradually adding to it. This mitigates project risk, addresses inevitable shifts in user needs and helps maintain information quality in the data warehouse. Finally, it will be ensured that the development project has a well-defined schedule and budget from day one – they work as important standards by which to measure project performance from the process perspective.

Technical considerations that stood out in the business case had to do primarily with security and documentation. It became clear how critical security was in all stages of data management,

and consequently it was seen how data availability can be seriously impeded as a result of highly restrictive access controls – this was highlighted with CaseCo in particular, given its sensitive industry. However, as concern for the privacy of individuals is today a hot topic in general, data security has quickly become important for businesses regardless of their industry. Businesses must then find a way to handle security concerns without crippling their data-driven decision-making in the process.

Further difficulties were encountered with the lack of documentation about the company's enterprise database, a primary data source for a new data warehouse. In the absence of proper information about data tables, columns and their formats and relations, the background work necessary for constructing a new data warehouse and its corresponding data pipelines took longer than it otherwise would have. That is to say, information that could have been gathered from documentation instead had to be transferred by staff members who were familiar with the database. This provided a simple but powerful demonstration about the importance of database documentation – while its benefits are not necessarily obvious, its lack can cause very disproportionate amounts of unnecessary work later down the line.

5.2 Limitations in research

Due to the relatively narrow scope that a BSc. thesis is characterized by, the literature review that was done is a cursory one, focusing mainly on those aspects of BI that seemed most relevant for a project like the one in which the researcher was employed. Likewise, the timeline for the thesis did not allow for the project's ultimate, long-term success to be assessed here.

In practice, the novelty of this research was also limited by the fact that the implementation described here is a relatively traditional one. Although traditional solutions were found suitable for this business case, it also became apparent during research that more and more companies these days should seriously consider a cloud solution for their BI back-end. Highly flexible thanks to on-demand up- and down-scaling and equipped with robust security features, the cloud offers an alternative that involves very little initial investment – and potentially less overall costs, too. These new dynamic architectures are applicable especially to business models that can make use of vast streams of "big data", be it from external sources like media outlets or internal ones like real-time process sensors.

REFERENCES

A.B. Sangar, N. B. A. Iahad,. 2013. Critical Factors That Affect The Success Of Business Intelligence Systems (BIS) Implementation In An Organization. International Journal of Scientific & Technology Research, Volume 2. p.176-178

Adamala, S. and Cidrin, L., 2011. Key success factors in business intelligence. Journal of Intelligence Studies in Business. 1.

Ariyachandra, T. and Watson, H., 2010. Key organizational factors in data warehouse architecture selection. Decision support systems, 49(2), p. 200-212.

Ang, J., Thompson, S.H.T., 2000. Management issues in data warehousing: insights from the Housing and Development Board. Decision Support Systems, Volume 29, Issue 1. p. 11-20.

Baskerville R.L. and Wood-Harper A.T., 1996. A critical perspective on action research as a method for information systems research, Journal of Information Technology, Vol. 11, No. 3, pp. 235-246

Bose, R., 2009. Advanced analytics: opportunities and challenges. Industrial Management & Data Systems, 109(2), pp.155-172.

Chan, J. O., 2005. Optimizing Data Warehousing Strategies, Communications of the IIMA, 5(1). p. 1-13.

Chandler, Neil., Hostmann, B., Rayner, N. & Herschel G,. 2011. Gartner's Business Analytics Framework

Cook, N.D., 2015. Crisis management strategy: Competition and change in modern enterprises. Routledge. p.10-28

Coughlan P. and Coghlan D., 2002. Action research for operations management, International Journal of Operations & Product Management, Vol. 22, No. 2, p. 220-240

Davenport, T.H. and Harris, J.G., 2007. The architecture of business intelligence. Competing on analytics: The new science of winning. Harvard Business Review Press.

DeLone, W.H. and McLean, E.R., 1992. Information Systems Success: The Quest for the Dependent Variable. Information Systems Research, vol. 3, no. 1, p. 60-95

DeLone, W.H. and McLean, E.R., 2002. Information systems success revisited. In Proceedings of the 35th Annual Hawaii International Conference on System Sciences p. 2966-2976.

Dobrev, K. and Hart, M., 2015. Benefits, justification and implementation planning of real-time business intelligence systems. Electronic Journal of Information Systems Evaluation, 18(2), p.106-112.

Fuchs, G., 2004. The vital BI maintenance process. Information Management, 14(7), p.12.

Gartner, 2013. Hunting and Harvesting in a Digital World: Insights From the 2013 Gartner CIO Agenda Report. [WWW-document]. [Accessed on 11.12.2019]. Available: https://www.gartner.com/imagesrv/cio/pdf/cio_agenda_insights2013.pdf

Gartner, 2019. 2019 CIO Agenda: Secure the Foundation for Digital Business. [WWW-document]. [Accessed on 11.12.2019]. Available: https://emtemp.gcom.cloud/ngw/globalassets/en/information-technology/documents/trends/gartner-2019-cio-agenda-key-takeaways.pdf

Idoine, C. & Howson, C., 2017. How to Enable Self-Service Analytics and Business Intelligence: Lessons From Gartner Award Finalists. Gartner Inc.

Kim, H., & Kankanhalli, A., 2009. Investigating User Resistance to Information Systems Implementation: A Status Quo Bias Perspective. MIS Quarterly, 33(3), p. 567-582.

Klie, L., 2015. Integrate CRM and ERP for better intelligence. Customer Relationship Management, 19, p. 28-32.

Marketwatch, 2019. Business Intelligence Market Is Booming at a CAGR of 26.98% by 2025 [WWW-document]. [Accessed on 10.10.2019]. Available:https://www.marketwatch.com/press-release/business-intelligence-market-is-booming-at-a-cagr-of-2698-by-2025-2019-01-09

Olszak, C.M. and Ziemba, E., 2007. Approach to building and implementing business intelligence systems. Interdisciplinary Journal of Information, Knowledge, and Management, 2(1), p.135-148.

Ong, In & Siew, Pei & Wong, Siew Fan & Rahman, Abdul & Malaysia,. 2011. A Five-Layered Business Intelligence Architecture. Communications of the IBIMA.

Popovicˇ, A., Hackney, R., Coelho, P.S., & Jaklicˇ, J., 2012. Towards business intelligence systems success: Effects of maturity and culture on analytical decision making. Decision Support Systems, 54(1), p. 729–739.

Project Management Institute, 2017. A guide to the project management body of knowledge (PMBOK guide). Newtown Square, Pa: Project Management Institute. p. 221-230

Turban, E., Sharda, R., Aronson, J.E. & David King,. 2010. Business Intelligence: A Managerial Approach. p.222-232

Watson, H., Fuller, C., Ariyachandra, T., 2004. Data warehouse governance: best practices at Blue Cross and Blue Shield of North Carolina. Decision Support Systems 2004, Vol.38(3), p.435-450

Watson, H. J., Annino, D. A., Wixom, B. H., K, L. A., & Rutherford, M., 2001. Current practices in data warehousing. Information Systems Management, 18(1), p. 47-55.

Williams, S., & Williams, N., 2007. The profit impact of business intelligence. Morgan Kaufmann Publishers.

Yeoh, W. and Koronios, A., 2010. Critical success factors for business intelligence systems. Journal of computer information systems, 50(3), p.23-32.

Yeoh, W. and Popovič, A., 2016. Extending the understanding of critical success factors for implementing business intelligence systems. Journal of the Association for Information Science and Technology, 67(1), p.134-147.