KARELIA UNIVERSITY OF APPLIED SCIENCES
Business & IT – VIVES BELGIUM

Cédric Durant & Florian De Langhe

GAINING VALUE OUT OF DATA – SMART CITY

Thesis
June 2015

THESIS
June 2015
Degree Programme in Business & IT

Tikakarinne 9
FIN 80220 JOENSUU
FINLAND
Tel. 358-13-260 600

Author(s)

Cédric Durant & Florian De Langhe

Title

Gaining value out of data – smart city

Abstract

This thesis is commissioned by Process Genius. Riku Saarenheimo was our mentor from the company and Petri Laitinen was our tutor from KARELIA. Process Genius visualizes industrial process plants for organizations.

The thesis is divided into two sections ; in the theoretical part we first describe big data, BI, data warehouse and data marts. The practical implementation of the project consists of two separate parts:

First we analyzed 2 USE CASES where we apply the theoretical background to create solutions for the city of Joensuu;

The second part, the practical application were we describe how we would implement the data marts into the data warehouse.

For completing this thesis we have used the theoretical knowledge and experience we have acquired during our bachelor's programme.

The main objective of this thesis is to conclude if big data, BI and a data warehouse can be used to develop a solution that improves the profit or reduce the costs of a city/company.

Language
English

Pages:109

Keywords

KARELIA University of Applied Sciences, Joensuu, Big Data, Data warehouse, Predictive analytics and Smart City.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

We are two students from VIVES Belgium making our final project as exchange students at Karelia University of Applied Sciences in Joensuu, Finland. We studied Applied Information Technology: Business & IT in Belgium. Thanks to our tutor Petri Laitinen we got in contact with the company, Process Genius. Process Genius gave us the task to develop a data warehouse and create the possibility to access data through a HTTP request.

Since we did not have knowledge to create a data warehouse or HTTP request, we changed our topic of the thesis by permission of Petri Laitinen to 'Knowledge needed to implement data warehouse and data marts and the use of it in Smart cities'. It enabled us to discuss about the theoretical aspects of a data warehouse and everything around it. Is Big Data really the future and can companies and even cities really gain value out of these big data BI solutions? We have tried to answer these questions in this thesis report.

The thesis report starts with the action plan. We will describe the companies we worked with, explain our project and the possible problems, and what the purpose of this thesis was. Then we present the theoretical background of our thesis. We begin with Big Data and explain why big data enables us to accomplish this project. Next we talk about business intelligence and the different business intelligence tools. The following chapter explains the data warehouse concept, how it works, and what the advantages and disadvantages are.

After this, we will talk about data marts, an important part of a data warehouse. Next we talk about predictive analysis, which uses big data, BI, data warehouse, and data marts

In conclusion, we will apply all this theory to create some use cases that provides solutions to make a city become a Smart City. We will also explain how we implemented the data marts into the data system that our colleagues Jonas Lesy and Ruben Vervaeke created for their final project.

1 ACTION PLAN

In this chapter we explain which organizations we work with, what they expect us to achieve, who will provide us with the necessary resources (hardware & software), and how everything will be done to accomplish our project. We also discuss the problems that can be encountered, the chosen approach of the project, and explain what we have done during our internship.

1.1 Project background

The project background contains a description of the different organizations with whom we worked during our internship. Also, our relation with them is explained to give a good understanding of the way of cooperation.

1.1.1 Organizations

During our internship, we contacted and worked with three different organizations. One of them, Process Genius, commissioned our project and they were the company for whom we with primarily collaborated. The second organization is the city of Joensuu. They provided us with the necessary data for our project. The last organization that helped us to achieve our project was Karelia University of Applied Sciences. Two tutors from the university guided and helped us during the time of our internship in Joensuu, Finland. They also provided us with the software and hardware to successfully end our project in the given time. Process Genius and the two tutors of Karelia University of Applied Sciences evaluated our project at the end of our internship.

1.1.2 Process Genius

Process Genius is a company from the city of Joensuu, Finland. They are specialized in online 3D services. Their services are primarily for industrial process plants and the surrounding sales companies of those industrial process plants. (Process Genius 2014)

Their services can be customized to be used in three different areas:

- PGtool: Is a tool that can be used to present a company's products in 3D within the processes of other companies. It can also be used to make a connection between the company's marketing network and the intelligent social media.
- PGplant: Is a package that is made for end users so that they are able to keep control over data, information, and documents that are generated by the company's process plant.
- PGedu: Is a tool that focuses on educational purposes and is used at educational and training institutions. The tool is still being programmed.

The main objectives of implementing their service into a company are:

- Improving productivity;
- Reducing costs;
- Saving time (in general).

Process Genius is a recently established company that started in 2011. The vision of the company is to offer their customers 'next gen' tools that have a variety of functions and that can be used for sales productivity and training. (Process Genius 2014)

The company was established by two experienced sales professionals and engineers that have a lot of experience in processes related to sales positions in companies. (Process Genius 2014)

Process Genius uses a team of qualified employees to achieve their vision. Each individual uses his know-how and skills to produce tools and products of high technology and high quality. They are supported by a group of advisors that have an impressive knowledge in finances and business material. (Process Genius 2014)

The following organization chart (Figure 1) represents the management structure of Process Genius:

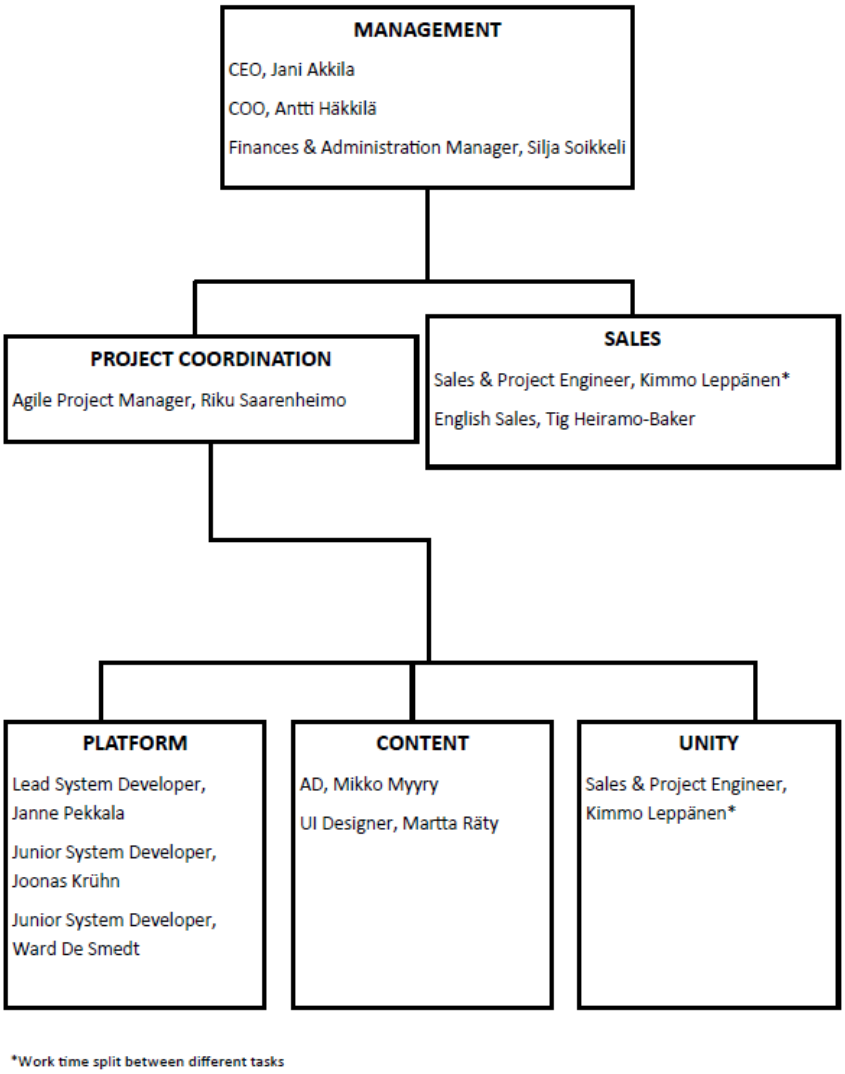

Figure 1 Hierarchical structure of Process Genius

1.1.3 Process Genius and us

Process Genius is the company for whom we worked during our internship in Joensuu, Finland. After a meeting with the CEO, Jani Akkila, and some of the employees of Project Genius, we received information about their company and what they want to achieve for their clients as a company. They also explained what our project was and what had to be realized to successfully achieve our project. At the end of our internship, Process Genius will use the information we have gathered to have a better understanding of how BI can be used in cities and will use the project result in their future projects for their clients. The exact description of our project is presented in detail later in this thesis report.

1.1.4 City of Joensuu

The city of Joensuu was our second partner and the one for whom we had to gather most of the information on Smart Cities and Business Intelligence. The city of Joensuu is located in the region of North Karelia in Finland. Joensuu counts a population of 75.086 habitants. The city is known for its high number of students, as it hosts two universities, the University of Eastern Finland and North Karelia University of Applied Sciences. (Joensuu City 2015)

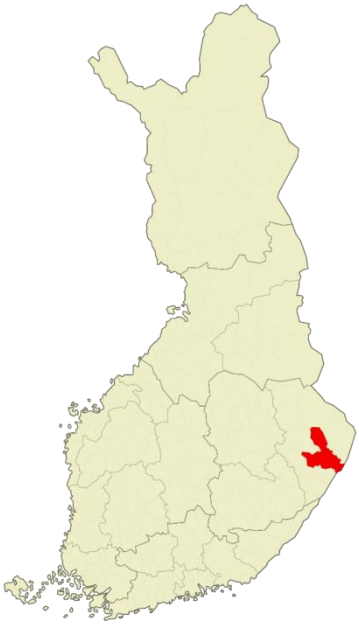

Figure 2 North-Karelia

Joensuu is the city where Process Genius is located and also where we stay during our internship. It is a city that tries to be ahead of other cities in matter of social, environmental, and economic aspects. They use new technologies to protect the environment and to resolve pollution issues that could affect their natural resources like forestry. Because of the persistent advance in technology, Joensuu has had quite many changes the last few years. Because of those changes, the city accumulates a lot of data coming from different installations and sensors such as, traffic lights, electricity usage, bus location, and bridge status. Most of the data is just gathered and saved without being fully analyzed. The wish of the city is to become a Smart City where all the gathered data is analyzed in order to fully enhance the different parts of the city. (Joensuu City 2015)

1.1.5 The city of Joensuu and us

The city of Joensuu provided us with the needed data for our project. The received data was first limited because of some technical issues. At the end of our project this problem was still not resolved because of the large amount of data that couldn't easily be sent. When this problem is resolved our project will be fully operational and they will be able to use it for future projects. Our main role here was to provide information about Smart Cities and Business Intelligence in general to the city of Joensuu and help them make decisions for later needs. Therefore, we examined some use-cases of other cities that have already implemented BI-tools, sensors, tracking devices, etc. to become a Smart City.

1.1.6 Karelia University of Applied Sciences

The Karelia University of Applied Sciences was the third partner that helped us during our internship here in Joensuu. The University is the second institute of higher education in Joensuu and gives students the opportunity to get a bachelor or master degree in different fields of study.

The degrees that can be achieved in the following fields:

- Culture;
- Social Sciences, Business and Administration;
- Natural Sciences;
- Technology, Communication and Transport;
- Natural Resources and the Environment;
- Social Services, Health and Sport;
- Tourism, Catering and Domestic Services.

The objective of Karelia University of Applied Sciences is to have students with high-qualified skills in order to provide the communities with highly qualified workers that will enhance the economy and help resolve the problems of today. For keeping a high level of education, the system is constantly reshaped to be in harmony with the needs of tomorrow. (Karelia University of Applied Sciences 2015)

Figure 3 Karelia logo

1.1.7 Karelia University of Applied Sciences and us

Karelia University of Applied Science is the institution that helped us accomplish our project. Two tutors from the university coordinated and helped us during the different stages. The university also provided us with various resources.

We got a room specially dedicated for our final project. The room was used to write our thesis, to do some research around the different subjects we describe in this thesis, and to do the practical part of our project. They also gave us the needed hardware. The hardware given by the University included servers located on Wärtsilä campus of the Karelia University of Applied Sciences. These servers were used to install the software needed for our project and to implement the final result that will be accessible by the city of Joensuu and Process Genius.

A good partnership between our coordinators and us was essential to achieve what was expected from us by Process Genius and the City of Joensuu. Their knowledge and experience guided us, and their different points of view on subjects could resolve some of the problems we encountered. The two coordinators will also be the one grading our final work and our thesis.

Every week there were meetings with our coordinators to discuss the progress we had made. These meetings did not only include the practical part, but also our progress in writing our thesis. They also scheduled the planning of our internship with us to be certain that everything is finished by the due date.

2 PROBLEM DESCRIPTION

The problem description is divided into an explanation of the project, the problems we could have encountered during the duration of our project and how we thought we were going to approach the project to achieve our goals. All of these aspects are explained in detail below because of their importance for the project.

2.1 Project explanation

Process Genius asked us to give them a way to process data into useful data and to help them acquire more knowledge about Business Intelligence and smart cities. The same information about Business Intelligence and smart cities was demanded by the city of Joensuu. The data that needed to be processed came from the city of Joensuu.

The main reason why Process Genius wanted to acquire this knowledge was to be able to use it later in projects for their clients and to expand their clients not only to companies, but also to cities like Joensuu. In this way they would have the possibility to add extra features to their solutions to help cities to become smart cities. Process Genius already has a partnership with the city of Joensuu to implement solutions that will process data coming from the city to a web service available for the habitants of Joensuu. They needed our aid for the data processing and the BI part.

The concept of a smart city is something the city of Joensuu is very interested in. They want to know how other cities have done it and what the results have been. In that way they can get an idea of how they can implement it themselves, what they all have to do, and which points they have to pay attention to. Our task was to provide them with information, use-cases and other methods needed to have a good view on how some solutions can affect their city.

2.1.1 Possible problems

The possible problems that we could encounter during our project will be listed here. This list gives a summary of things that we had to keep an eye on. If any of these problems were not resolved at an early stage or are not prevented to begin with, they could have delayed the project or prevented the project to be fully finalized as intended. Every problem that we present is also divided into three different categories. Each category is in regard to the potential encounter of the problem. The three used categories are low risk, medium risk and high risk.

List of possible problems we could have encountered during our project by categories:

- Extension of the scope;

This problem was graded as a low risk because we do had a good understanding of what Process Genius and the city of Joensuu expected from us. We also discussed with our coordinators what should and should not be examined, this to have a delimited scope.

- Availability/Accessibility of the client;

We also placed this problem in the low risk category because we could easily contact Process Genius by mail, Facebook, or phone. We were also always welcomed at their office if we had questions or needed help to achieve a task.

- The language barrier;

At first we thought that this problem was going to be classified in the medium risk category. However, after the first meeting with our coordinators and the people of Process Genius, we realized that our English skills were good enough to communicate effectively. Being able to understand each other is the key to success and that is why it was important that we had the same level of English communication skills. So this problem was categorized as that of low risk.

- Lack of knowledge and information about the implementation of data marts;

Considering we had never done similar projects before, our knowledge the subject was inexistent. We knew the theory behind data marts so we hoped that it would help us in the implementation. Internet is a good source for information but it did not mean we were able to find everything we were looking for. This is why this fact was classified as a medium risk problem.

- Lack of information on the desired subjects;

Since most of the subjects we discuss in our thesis are quite new and still developing, it was possible that we would not have been able to find enough information about them in books or on the Internet. If this is the case, we had to discuss with our coordinators how to find relevant information for our thesis and still be able to accomplish what Process Genius and the city of Joensuu expect of us. This problem was classified as a medium risk problem.

- Lack of time;

Time management is very important when realizing a project. Everything needs to be scheduled and deadlines have to be kept to be able to finalize everything in time. If something takes more time than anticipated, it could affect the rest of the project and delay it. However, because our internship was restricted in time we would not have another chance to reschedule things if something was to go wrong. That is why this problem was classified as a high risk problem.

- Inadequate management of the project;

From the start we had to discuss how everything needs to be organized for the project: Who will do what and how it needs to be done? The management of a project can easily go wrong without recognizing it. The problems often start to

appear only near the end of the project and then it would be too late to make changes. This problem was classified as a high risk problem.

- Data issues;

For processing data into useful information, many steps needed to be taken in the right order. The data needed to be cleaned and shaped in the right form before we even had the possibility to do something with it. If the steps were not done right, or the data was too polluted, it could have resulted in a big problem for us to accomplish the data processing. Because data was very important for our project this problem was also placed in the high risk category class.

- Software issues;

During our project we needed some software to process the data and to create the data marts. Software can easily cause some issues when it is not installed or used as it should. This is why we had to control everything several times to avoid errors, bugs, and other software problems. This was classified as a high risk problem.

- Hardware issues;

Hardware issues are the most problematic because often hardware components are expensive to change or replace. It was also possible that the hardware provided by Karelia University of Applied Sciences would not be powerful enough for our project and therefore we would not be able to achieve the practical part of our project. This is why this problem was classified as a high risk problem.

2.2 Project approach and goals

In this section, we discuss how we approached the project, what our goals were and which tools and which hardware we used to accomplish everything.

The project was divided in two parts. The first and main part included research and the second part was the practice based project. For the first part, we researched different subjects using print and online materials.

Here is a list of subjects that were researched:

- Data;
- Big Data;
- Business intelligence;
- BI-tools;
- Data warehouses;
- Data marts;
- Star schema;
- Snowflake schema;
- Predictive analytics;
- Smart cities;
- Use cases.

Our first aim was to successfully give Process Genius and the city of Joensuu information about those subjects in a comprehensive format, so that they could use the new acquired knowledge in their future projects.

With the information gathered about these subjects, we already had the needed information that was requested by Process Genius and the city of Joensuu. The information that we collected was also useful for us in order to explain the use-cases that we found and that were relevant and of important enough to be added to our thesis. The use-cases were fictive or from other cities that had already implemented these tools to become smart cities.

The second goal consisted of the successfully implementation of data marts in a data system. This goal was not the main objective so we first focused on our main goal and if we still had enough time we would start with the implementation of data marts using the information we will had gathered to achieve our first goal.

The practical part included the implementation of the data marts and the use of Business Intelligence tools to process the data received from the city of Joensuu into information that can be used. This part will only be realized after Jonas Lezy and Ruben Vervaeke had implemented their part of their project because our data marts are linked to their data warehouse and their web service.

For the data marts we planned to use Hive. The data marts were to run on the servers provided by Karelia University of Applied Sciences. The Business Intelligence tools that have to be used still needs to be discussed. This part of the project isn't described in the thesis because it wasn't done before the due date of the thesis.

The whole scheduling of the project was placed on Trello. This is a program that allows to easily follow the progress of a project and to schedule every part of it by using deadlines and milestones.

3 DATA & BIG DATA

Data is information generated by certain events inside a company/city. In these modern days, a lot of data is generated without us knowing and is saved in large datasets. These dataset are often so large that it is impossible to process using the traditional processing methods.

When this occurs it is called big data. McKinsey articulated the following definition of big data in his 2011 big data study:

“Datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.”[1] This definition is intentionally subjective because the definition of how big a dataset needs to be is fixed. This is why we first have to define the three V’s.

The three Vs were invented by the industry analyst Doug Laney[2]. He wanted to define how big the dataset has to be to be called big data. The three Vs are :

- Volume; the size of the data has to be big enough to be called Big Data. It is not only the size (Gigabyte, Terabyte) that matters but also the value and potential of the data. In general the more data you have, the higher the accuracy is of the predictions. (Thomas H. Davenport Jill Dyché 2013)
- Velocity; refers to the speed of generation of the data. How fast the company/city generates the data. If the company needs real time data it is necessary that the velocity rate is within seconds. (Thomas H. Davenport Jill Dyché 2013)
- Variety; all the different types of formats inside the datasets. For example, structured data, unstructured data, documents, video, etc. (Thomas H. Davenport Jill Dyché 2013)

[1] http://www.forbes.com/sites/davidwilliams/2012/09/19/if-big-data-simply-meant-lots-of-data-we-would-call-it-lots-of-data/

[2] http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

When every V is high, we then can call the dataset Big Data. Before we go into detail what Big Data can mean for companies and cities, we would like to explain the history of Big Data. The history is important to know so then we can understand where it comes from and how it got to this stage.

3.1 History of Big Data

At the end of the 1900's, several scientist discovered that the growth rate in the volume of data is exponential instead of linear, which is called the "Law of exponential increase"[3]. This means that approximately every two years the data volume will be doubled, also known as the information explosion.

This made the scientists wonder about what will happen in 10 years and if we will still be able to collect all the data.

In 2000, Francis X. Diebold presented to the Eighth World Congress of the Econometric Society a paper titled "Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting"[4] In this paper, Francis talked about the Big Data phenomenon. This made a lot of scientists realize how much they can benefit from Big Data. The term Big Data was born. (Gil press 2013)

One year later in 2001, Doug Laney published a research note titled, "3D Data Management: Controlling Data Volume, Velocity and Variety"[5]. This paper defined the foundation of Big Data. To this day people still refer to the 3Vs of Big Data when explaining Big Data. (Gil press 2013)

Randal E. Bryant, Randy H. Katz and Edward D. Lazowska published in 2008 „"Big – Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society"[6]. In this paper they talked about the advantages of using Big Data inside the Company and how it will affect the activities and processes.

---

[3] http://en.wikipedia.org/wiki/Moore%27s_law, http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

[4] http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

[5] http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

[6] http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

They talked about how Big – data is perhaps the biggest innovation in computing the last decade and that we only begun to see its potential to collect, organize, and process data. This is still true today. (Gil press 2013)

In May 2012, Danah Boyd and Kate Crawford published “Critical Questions for Big Data”[7]. They defined big data as “a cultural, technological and scholarly phenomenon that rests on the interplay of Technology, Analysis and Mythology”[8]. What they meant with this is that Big Data is only possible thanks technology and the maximization of computing power and algorithmic accuracy to gather and analyze large data sets. Analysis and the possibility to find patterns in these large data sets and Mythology to gather knowledge that was previously impossible. (Gil press 2013)

By now, every company understands the importance of Big Data but a lot of companies fail to analyze this data to their advantage, or a lot of companies/Cities don’t know how to start collecting this data and analyze these huge data sets.

We will try to explain how to tackle this problem in the next Chapters.

3.2 Unstructured and structured data

When we are talking about Big Data we have to understand that Big Data is a collection of data from traditional and digital sources. These sources can be inside and outside the company. The problem with Big Data is that some people only want to collect the digital sources since it is easy to collect and store this type of data. But it is important to also include the traditional sources. This causes a problem since the data from the digital sources are unstructured and the traditional data is structured.

Unstructured data is information that is not organized and can’t be stored into traditional databases or data models straight away. Examples of unstructured data are tweets, metadata, and social media posts. (Lisa Arthur 2013)

[7] http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

[8] http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/

Structured data is data stored according to a certain data model. A data model defines how the data is stored, processed, and accessed. (Lisa Arthur 2013)

The challenge with Big Data is to get both the unstructured and structured data into the same database. This is possible thanks to a Data warehouse which we will explain in a different chapter. It may seem that Big Data is all rainbow and sunshine but in reality this isn't the case. There are still a lot of problems with Big Data.

3.3 Problems of Big Data

The 8 problems we will discuss are:

- Budget
- The know-how of the IT
- The arrangement of the Data
- A mountain of data
- New data centers
- Storage of data
- Big Data sellers
- Aligning business and IT

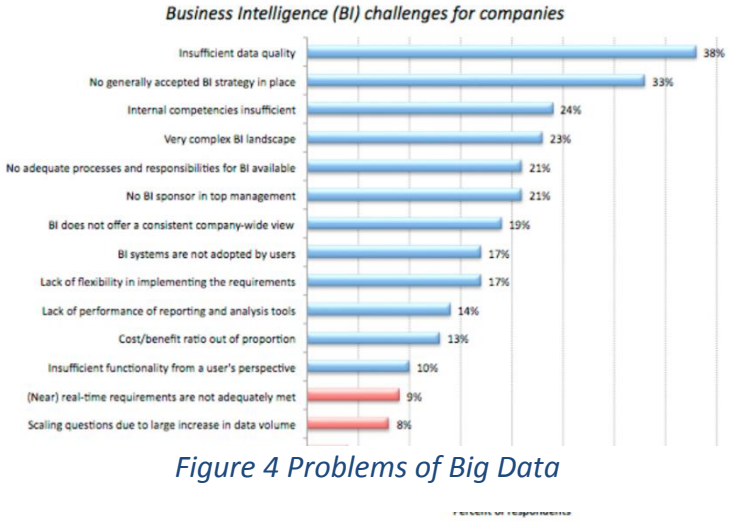

Figure 4 Problems of Big Data

3.3.1 Budget

A lot of servers inside data centers aren't made for analyzing and processing of Big Data. When you want to process and analyze Big Data the company needs strong and powerful servers and applications. This means that if the company wants to get started with Big Data they will have to invest in new IT solutions. A lot of companies simply don't have the money for this and even if they had, it would be hard to convince the board members to invest in new equipment while they still have in their eyes perfectly working old servers. The CIO (Chief Information officer) will play an important role to convince the board members.

3.3.2 The know-how of the IT

The processing of Big Data is completely different as normal data. It requires another way of approach and therefore, a different approach in terms of processing and storage. Some companies don't respect this and this causes the processing and storage to be chaotic and not productive. As a results this causes a lot of problems for the companies.

The best way to store the data is according by the importance of the data. Data that is important needs to be stored on quick accessible storage devices. This way the data can be quickly consulted and visualized. The storage devices are more expensive than normal storage devices, therefore we have to use these storage devices is the most cost effective way. The data that is important needs to be stored on these storage devices, while all other data can be stored on slower and cheaper storage devices.

3.3.3 The arrangement of the data

Big Data and analysts are very dependent on data. If this Data is dirty or inaccurate it will still be dirty/inaccurate after the processing of the data and the analysis will also be wrong. This way the "cleaning" of the data is so important. Dirty data has to be deleted before a company can start a Big Data project. The cleaning part asks a lot of time and money. This is again hard to accept for the board members because they only see how much money it costs and see no clear result. It is the task of the Big Data sellers or IT department to explain why this is an essential part of a Big Data project.

3.3.4 A mountain of data

The amount of data is almost quintupled in comparison with 3 years ago. This means that the companies have a lot more data to be processed and analysed. The problem is that in terms of managing all this data there isn't a lot of improvement in comparison with 3 years ago. This means that in order to process and analyse the data, the companies have to find a "key" and this is only possible when the data is arranged. This shows the importance of the arrangement of data and how big of a task it actually is to process huge amount of data.

3.3.5 New data centers

The old data centers are made for transaction processes and this means that during the night or when the servers aren't overloaded, a new batch will be loaded into the servers. Because of the upswing of Big Data, business analysis companies want to run these analyses on real time data. This causes the data centers to change their setup to be able to provide real time data.

3.3.6 The storage of data

The fear of many companies is to lose their data. That is why is it so important to store the data in the right place in the right way. Back-ups are necessary because IT infrastructures and systems aren't 100% trustworthy as of yet.

3.3.7 The Big Data sellers

The Big Data sellers know that a lot of companies are still inexperienced with Big Data. This is way they try to sell premade systems that only needs to be implemented inside the company. This is good way to start with Big Data but after a while the companies needs systems that are custom made for their company. When choosing a Big Data seller it is smart to choose one where the system is adaptable.

3.3.8 Aligning business and IT

Before investing into IT, the companies have to make sure that their goals are matched with the Big Data strategy. They also have to look if the implementation of the new systems and infrastructure go against the goals their business goals. There is also need to study what the profit and costs will be when using Big Data. If there are more costs than profits it would be better for the company to not invest in their IT. They also have to look whether they will use the Big Data, because sometimes it is better to keep using the old system.

4 BUSINESS INTELLIGENCE

Business intelligence or BI, is often always heard together with Big Data. But what is Business intelligence? We will explain BI in this chapter.

BI can be defined in three ways:

- BI as a process
- BI as a technology
- BI as a phenomenon

4.1 BI as a process

Business Intelligence is a “continuous process that provides companies to gather data, analyze and use this to improve decision processes to enhance the performance of the company.”[9]

What this means is that business intelligence will gather and analyse data in a way that is aimed to support the management. This process of gathering data and analyzing it will be explained in its own chapter. (Daan van Beek 2014)

4.2 BI as a technology

Business Intelligence is a “collection of ICT tools that supports and improves the BI as a process and it gives BI as a process a face.”

There are a lot of ICT tools that support and improve the BI as a process, which will be discussed in the chapter “Business Intelligence tools”.[10] What we mean with “it gives BI as a process a face” is that thanks to the tools we can see graphs and trends that make it easier to understand to someone who has never heard of BI before. If these tools didn’t exist, it would be a lot harder for someone that does not know about the whole BI world to understand why the company actually uses BI. (Daan van Beek 2014)

[9] https://www.biaward.nl/wat-is-business-intelligence-bi/

[10] https://www.biaward.nl/wat-is-business-intelligence-bi/

4.3 BI as a phenomenon

Business Intelligence is “a collection of strategies, concepts, culture, structure, standards and ICT tools that allow the companies to develop and be intelligent.”[11]

What this means is that Business Intelligence is not fixed. It is not a set of rules you have to follow and implement. Each organization may have their own view on how to implement BI. They decide how to use it in order to gain an advantage and what tools they use to gather and analyze all their data. (Daan van Beek 2014)

Of course there are certain standards and structures inside Business Intelligence but there is enough freedom to give it your own touch.

4.4 Business intelligence process

The process of implementing BI in a company has three phases:

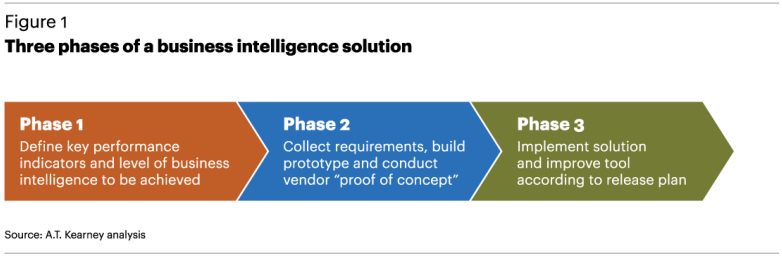

Figure 5 Three phases of BI solution

4.4.1 Phase 1

Define key performance indicators and level of business intelligence to be achieved. The first thing you have to do as a company is to define the KPI’s. A KPI has 3 elements:

- Objectives, these are the strategic targets transformed into operational targets.
- Measures, this is used to compare the vision with the strategy.
- Targets, is a goal and the review of the KPI is based on that goal.

[11] https://www.biaward.nl/wat-is-business-intelligence-bi/

We have to be careful when choosing the KPI's. A KPI has to reflect the company's objective because a wrongfully chosen KPI can give wrong results which can influence the decision making in a bad way. (Jenkel, R Simons, E Martin, A 2014)

When all the KPI's are defined, the company has to look if the BI system is only needed for the KPI or does it need to also include customer relationship management or supply chain management? Will the project be integrated into other departments/operations? When improving the finance BI, wouldn't it be good to also improve the Human relationship management (HRM) and customer relationship management (CRM) since they are connected with each other inside the company? When all these questions are answered we can start to look for the right BI tool. This last part of phase 1 and phase 2 is very important. Since a BI solution is very expensive it is necessary to choose the right tool straight away instead of using and modifying the wrong tool for years and eventually have to switch to another tool and wasting a lot of IT resources. (Jenkel, R Simons, E Martin, A 2014)

4.4.2 Phase 2

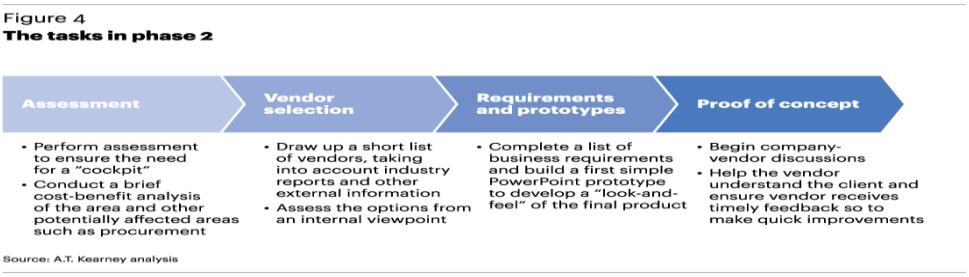

Figure 6 Phase 2 of BI solution

Assessment:
This is the most important step of the Business Intelligence tool, gathering all the right information about the situation now and the situation you want. The way a company should do this is through the following scheme.

AS Built	TO BE Built
AS IS	TO BE

AS IS
The current situation; current processes and workflows. The vision and strategy of the company and their values.

TO BE
What is the company aiming for when implementing the BI tool? What needs to be optimized, what is the scope of the project? What does the company want to accomplish when using the BI tool?

AS Built
What hardware and software does the company have at the moment.

TO BE Built
What hardware and software does the company want/need in the future? This is based on all the previous steps.

All these steps are in order AS IS > TO BE > AS Built > TO Be Built. It is essential that these steps are followed in this order. We can only look at the future when we know the present.

Vendor selection
When the company know all this we can start to look for a vendor that will help develop the BI solution. A lot of factors play into choosing the vendor. But for us, it is important that the vendor has a lot of experience because creating a BI tool isn't only looking at the facts but also "feeling the company" and looking at how the company works and values in order to make sure the BI tool is in line with the philosophy of the company. (Jenkel, R Simons, E Martin, A 2014)

Requirements and prototypes
When the company chooses a vendor, they work together to develop a prototype. It is important to always make a prototype first. When creating and testing the prototypes, the company always discovers some problems/things they didn't think about when brainstorming about the assessment. If the company doesn't

develop a prototype and develops the tool straight away and discovers these problems after the tool is implemented, it will cost the company a lot of money to try and fix these problems. It can even be possible to need to start over from scratch because the problem is too big to solve. This is why it is important to create this prototype. (Jenkel, R Simons, E Martin, A 2014)

Another reason why it is useful to create a prototype is to familiarize your employees with it so they get accustomed to it when implementing it. We call this user adoption. It is important that the employees want to work with the new tool. If not, the tool is useless because it doesn't get the data that's needed or it gets the wrong data from the employees.

Proof of concept

This is when the tool is developed and actually implemented in the company. It is important that the company works closely together with the vendor during this process. Sometimes it is hard for the vendor to understand what the company wants and if the company works closely together and gives timely feedback it is a lot easier for the vendor to create a tool the company really wants and will be accepted by the employees.

4.5 Phase 3

The is the last phase of the business intelligence process. The implementation is a project for the IT department. The implementation of the BI tool happens with the use of a release plan.

A release plan makes sure that the results are delivered quickly and the tool is implemented in the fastest and correct way. It divided the implementation itself into different parts. The first step is to implement all the dashboard functions and reporting requirements. It basically implements the structure without the data set. When this is implemented it is necessary to cleanse all the existing data and create a process that captures the new data in the right way. When the structure is in place and the data is captured in the proper way, the dashboard and reporting functions needs to be fine-tuned so it is easy to use for the management. The KPI's can either be implemented from the start or added after all this is done. (Jenkel, R Simons, E Martin, A 2014)

When all the phases are completed the company has a BI tool that is made for their company and hopefully will deliver the data that the company needs to make decisions. Again it is important that these steps are followed as it will cost the company a lot of money if a BI tool is wrongfully implemented or worst case not suitable for the company.

4.6 The advantages of Business Intelligence

4.6.1 Improved decision making

When we analyze a Business Intelligence solutions or tool, we look at the value it adds to the company and what the company benefits from it. In all the cases Business Intelligence needs to improve the decision making and analysis and reporting of the company. Business Intelligence improves the decision making thanks to 4 elements:

- "Required information is available;
- Data is consistent across organizational units;
- Information can be easily analyzed;
- Reports are presented in a use friendly format"[12]

These 4 elements are helping the company tremendously in their decision-making. When all the necessary data is available and the data is consistent and can be analyzed, the results are correct. And when these results are clearly presented in a report it is very easy for a company to understand it and take it into consideration when making a decision. (Jenkel, R Simons, E Martin, A 2014)

4.6.2 Efficiency on both IT and business side

On the IT side, the BI takes over the task of creating and changing data reports because the end-users of the BI tool can create and change their own reports.

[12] http://www.atkearney.com/paper/-/asset_publisher/dVxv4Hz2h8bS/content/using-proper-business-intelligence/10192

On the business side, the BI does the analyzing of data in a much faster and efficient way as the data from the reports come directly from the BI tools. This has as an advantage that the reports are a lot more credible as there are almost no human errors in play and it is also a lot more up to date because these reports can be remade in a matter of minutes.

5 BUSINESS INTELLIGENCE TOOLS

Business Intelligence tools are designed to “retrieve, analyze, transform and report data for business intelligence”[13] There are 5 types of business intelligence tools:

- Spreadsheets
- Reporting and querying software
- OLAP
- Digital dashboards
- Data mining

5.1 Spreadsheets

A spreadsheet is an “interactive computer application program for organization, analysis and storage of data in tabular form”. What this means is that a spreadsheet is in a certain way a database with some analyzing tools added. The spreadsheet has data in tabular form and can analyze this data thanks to formula’s that enable the user to calculate certain values. It also has rows and columns, which makes ordering of data a lot easier. (Wikipedia community 2014)

The best example of a spreadsheet is Microsoft Excel.

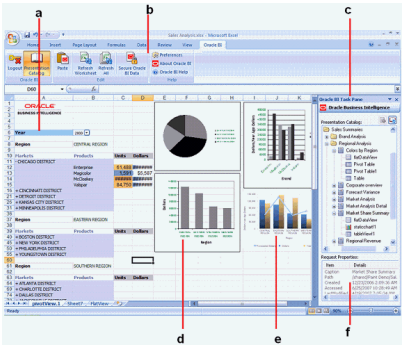

Figure 7 Spreadsheet

This is an example (figure 7) of a spreadsheet made by oracle. Oracle is one of the leaders of Business Intelligence tools.

In this example you can see that you are able the visualize data through diagrams (d, e).

It is also possible to make a pivot table. Thanks to this the user can easily and in a dynamical way summarize, order, group and edit the data. A pivot table can also to mathematical functions like calculate the average, maximum, minimum, etc. It can also

[13] http://en.wikipedia.org/wiki/Business_intelligence_tools

divide the data into different levels. For example all the data from year 2013, in the area Joensuu, from the Karelia school, class 15B.

5.1.1 Advantages of a spreadsheet

- Easy to use once you get to know the basics;
- Graphs will be automatically changed when changing a value in the data source;
- Calculations will be automatically changed and recalculated when changing a value in the data source;
- No additional software is required, the company just needs to install the spreadsheets software;
- Guidance/explanation is easily added through the use of notes. This makes a spreadsheet very user friendly.

5.1.2 Disadvantages of a spreadsheet

- Redundancy, each user has their own but the same data source.
- The changes made in the data source are locally. Each user works with its own version.
- The storage space is limited. There are only a few thousand rows available to store values into.

5.2 Reporting and querying software

Reporting and querying software helps the user to create reports, listings using queries to gather and analyse the data.

5.3 Query tools

The query tools are BI tools that are using queries to gather information and analyse data from the database. These queries enable the analyst to do sorting, filtering, searching and much more advanced methods. (Thomas C. Hammergren 2014)

There are a lot of query tools both paid and open source. It all comes down to the same basics so we will discuss one of the many query tools. The tool is called “AQT” or Advanced Query Tool.

The tool has 7 features:

- Database Explorer
- Run SQL
- Displaying Data
- Updating Data
- Graphical Query Builder
- Data Export
- Charting

These features are almost always in every query tool. Some tools might have more features and others might have less but in general these are the 7 main features of a query tool. Because of this we will explain all the 7 features.

5.3.1 Database Explorer

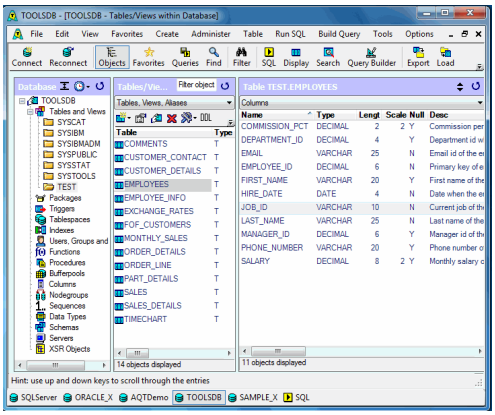

Figure 8 Database Explorer

The Database Explorer, it can also have a different name in other tools but we use this name because that’s the name AQT gave it.

The Database Explorer shows a lot of information about your Database. It shows every object, views, table, rows, columns, etc. (Cardett Associates 2013)

It also shows the database scheme, the dependencies between tables, and primary and foreign keys.

This helps the user understand their database and if necessary enables them to change/add a table, value or relation.

5.3.2 Run SQL

This is the core of a query tool. In this part of the query tool, the queries are made that enable the company to do filtering, sorting, searching and other methods

This Run SQL feature has several functionalities:

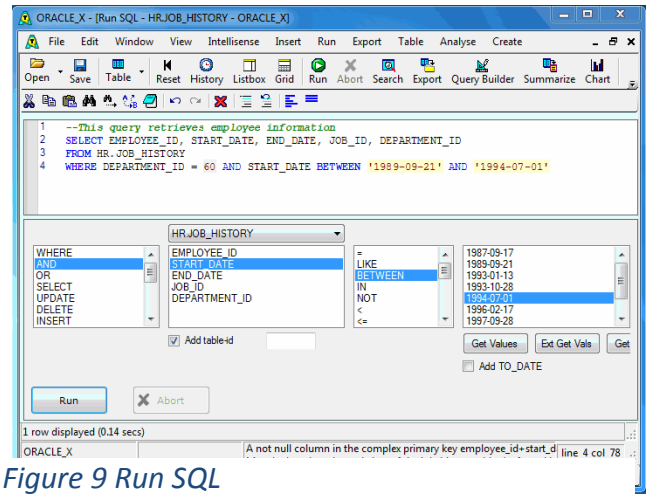

Figure 9 Run SQL

- Save and retrieve queries
- Export data
- Run SQL commands
- Compose SQL commands
- Syntax check of the SQL commands

This is basically the heart of the tool. In this feature the whole functionality of the tool is made by the queries that are being made by the users. It is here that the queries are built to sort, filter, search, analyze. (Cardett Associates 2013)

It is possible to run the SQL against your database straight away by pressing the Run button. This is useful for the SQL writer whose job is to compose the right SQL statements into a command that delivers the right result. (Cardett Associates 2013)

It is also possible to make and run multiple scripts containing thousands of SQL statements. The scripts can be paused, amended or restarted at any possible moment. (Cardett Associates 2013)

5.3.3 Display data

This window shows the table contents or generated data form the scripts in an understandable format.

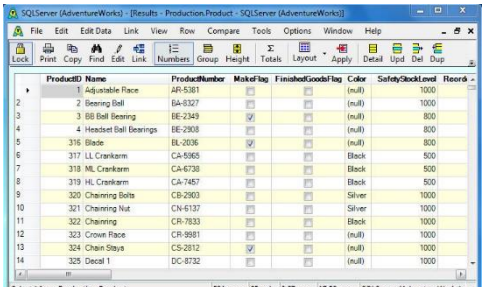

Figure 10 Display data

The data in the grid can be printed, sorted, copied, exported to excel or saved.

It is also possible to move the columns around or hide them. (Cardett Associates 2013)

There is also an option to format the display. For example, the results can be grouped by column color, subtotals can be calculated, and different styles can be applied to the data either build-in or custom made. (Cardett Associates 2013)

5.3.4 Updating data

This feature handles the update/delete/ insert SQL statement. Since these statements can ruin a whole database in just one second it is important to build in some fail saves when using the statements. The updating data is the fail save the tool needs.

When updating/deleting the tool shows how many rows are affected by the update. If this number is higher than the SQL writer expected it means that there might be something wrong with the statement. It is also possible to enable the "transaction" mode which will cause the user to explicitly do a Commit when changing the data in the database.

5.3.5 Graphical query builder

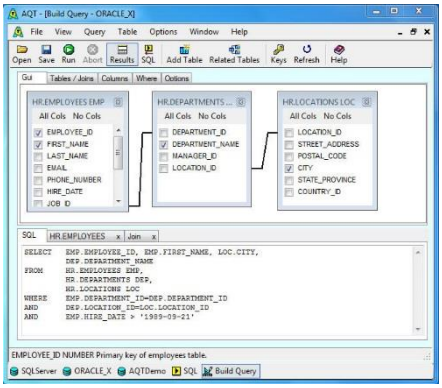

Figure 11 Example query builder

This feature makes the writing of SQL commands much easier. Thanks to the GUI, queries can be built much easier and smoother. The graphical query builder will do half of the work automatically. This query builder can build everything the Run SQL can, but it is much more user friendly. (Cardett Associates 2013)

5.3.6 Data export

This feature enables the user to export their results to csv, HTML, Excel,…

5.4 Reporting tools

These query tools are often hard to read. This is why reporting tools are invented. These tools make it so the results of the queries are understandable. The reporting tools translate the query results into a generated report. This report is understandable for the management level to take into consideration when making decisions.

5.5 OLAP

Stands for Online Analytical Processing. The result we get from OLAP is a multidimensional cube. The aim of the cube is to optimize the company and is used for reporting, analyzing, modelling. The implementation of OLAP is quick and easy to change. OLAP can be used in different “forms”

5.6 MOLAP

Is the more traditional way to OLAP analyzing. In MOLAP the data is stored into a multidimensional cube. The storage isn’t in a relational database but in priority formats

5.6.1 The advantages of MOLAP

- Excellent performance: MOLAP cubes are built to gather data and are optimal for “slice and dice operations.
- MOLAP can handle complex calculations. All the calculations are generated before the cube is built. Therefore the complex calculations are not only … but can be showed quickly.

5.6.2 The disadvantages of MOLAP

- It can only storage a limited amount of data.
- When using MOLAP there will be upgrades to the IT infrastructure needed.

5.7 ROLAP

ROLAP is a methodology based on manipulating the data that is stored in the relational database. The data is manipulated to give them the look of the traditional ‘slice and dice” functionality. This basically means that ROLAP adds a “where” clause in every SQL instruction.

5.7.1 The advantages of ROLAP

- It can store big amount of data. The amount depends on the amount of data in the underlying relational database. This basically means that ROLAP has no restrictions.
- ROLAP can use the functionalities that are inherited to the relational databases.

5.7.2 The disadvantages of ROLAP

- ROLAP is slow because every ROLAP report is essentially one or more SQL queries. The bigger the amount of data the longer it needs to process the queries.
- ROLAP is limited to the SQL functionalities. When it isn’t possible with SQL, it isn’t possible with ROLAP.

5.8 HOLAP

HOLAP technology tried to combine the advantages of MOLAP and ROLAP into one (HOLAP). HOLAP uses the cube technology of MOLAP quickly to do a “to do” task. But when specific information is needed, HOLAP uses the ROLAP technology to drill through the cube to consult the data in the underlying relational database.

5.8.1 The advantages of HOLAP

- HOLAP can handle big amount of data
- The cubes that HOLAP uses are smaller than the MOLAP ones because HOLAP can collect the detail data in relational database

- HOLAP is quicker than MOLAP because only aggregations are placed in the multidimensional format
- Long wait times because processes are only performed when changes are made.

5.8.2 The disadvantages of HOLAP

- HOLAP can be as slow as ROLAP for certain tasks
- HOLAP is obligated to perform when new records are added

5.9 Digital dashboards

“A digital dashboard is a visual display of most information needed to achieve one or more objectives which fits entirely on a single computer screen so it can be monitored at a glance”[14]

What this means is that a dashboard is a visual display of the most important information that is needed to achieve an objective. A dashboard can be fully visualized on one screen so it can be monitored at a glance.

[14] http://www.dashboardinsight.com/articles/digital-dashboards/fundamentals/what-is-a-dashboard.aspx

Example.

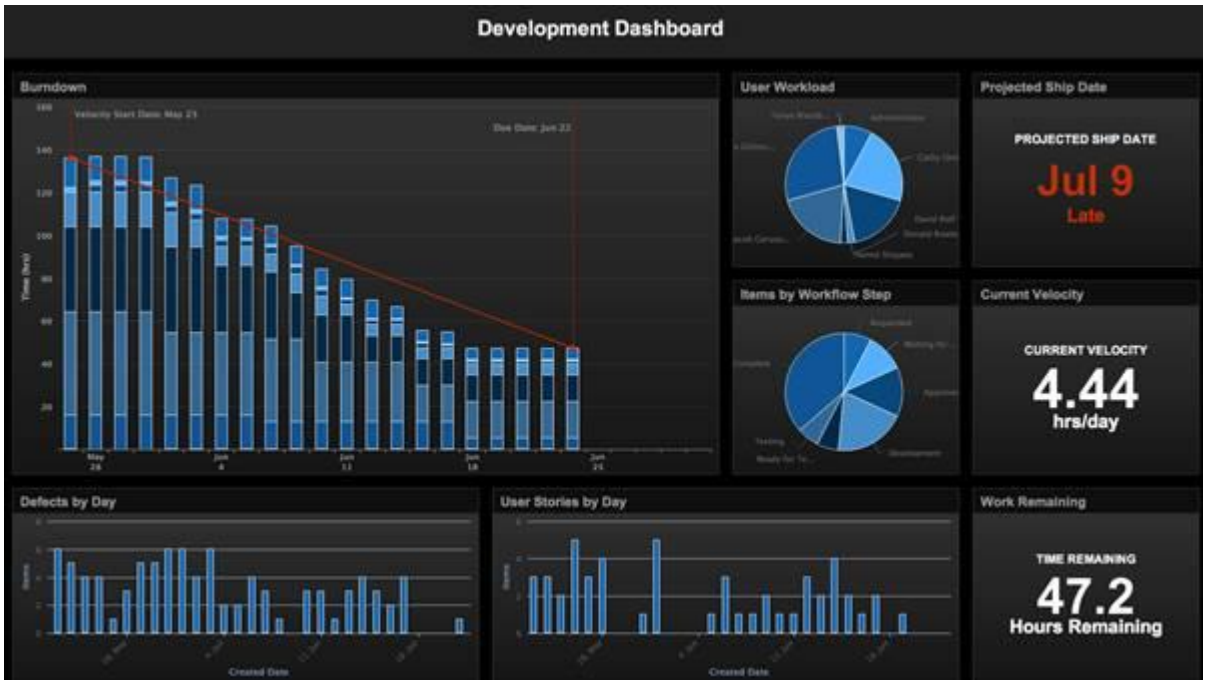

Figure 12 Example dashboard

This example is a dashboard that visualized a sprint. The dashboard summarized the data in a way so that it shows all the necessary data to follow the sprint on one screen. It also tells what happens during the sprint. For example, right now the sprint will be late, but it doesn't tell why it will be late. It gives empowerment to the user.

5.9.1 Empowerment

Empowerment allows the operational decisions to be made on the most optimal level. In this way the right people can take the right decisions based on the dashboard.

5.9.2 Advantages of empowerment

- The possibility to detect new trends;
- Reduction of costs;
- Decision making is based on facts;

5.9.3 Disadvantages of empowerment

- It can possibly work demoralizing;
- Missing the whole picture because we are stuck looking at the dash-board.

6 DATA WAREHOUSING

We find a data warehouse the ideal solution to implement Big Data in a company and because almost every multinational has a data warehouse we conclude that in our opinion a DWH is essential within the big data world. But what is a data warehouse?

There are infinite amounts of data warehouse definitions. But inside the world of data warehouses there are 2 pioneers with each their own definition:

- Ralph Kimball, “A DWH is a copy of transaction data specifically structured for query and analysis.”[15]
- Bill Inmon, “A DWH is a subject oriented, integrated, time variant and non-volatile collection of data designed to support the decision making process”[16]

Bill Inmon is the father of DWH, he was the first to define a DWH and years later Kimball tried to change the Big Data world with his definition which was actually nicely countered by Inmon. The reaction of Inmon on the new definition was “you can catch all the minnows in the ocean and stack them together and they still do not make a whale”[17] With this Inmon meant that a DWH has to be designed from top-down to include every data inside the company and only after this you can create data marts which we will explain in one of the next chapters. (Abramson, I 2013)

We associate us with the definition of Bill Inmon and therefore we are going to explain the definition in detail:

- Subject oriented, this means that the data is saved around severable subjects and not around certain departments.
- Integrated, the data is extracted from several sources and put into one database.

[15] http://www.1keydata.com/datawarehousing/data-warehouse-definition.html

[16] http://www.1keydata.com/datawarehousing/data-warehouse-definition.html

[17] http://ioug.itconvergence.com/pls/htmldb/DWBISIG.download_my_file?p_file=2346.

- Time variant, the DWH saves snapshots of the database to make a historical overview.
- Non-volatile, the DWH will not allow any changes of the data inside the database. If there are changes the change data will be saved as new data.
- Collection of data, a DWH is a database with a collection of data from different sources and systems like a OLTP-system.
- Designed to support, the DWH is designed to support the user in a performant way so he can easily research the data.
- Decision making process, The DWH is designed so it can be uses by BI-tools who support the decision-making process of the management.

(Abramson, I 2013)

Now that we have a clear view of the data warehouse we can discuss which kind of data are playing what role inside a DWH.

There are 4 kinds of data inside a DWH:

- Current detail data, these are very detailed data inside the DWH. This data is used when an analyst needs to analyze something very specific.
- Summarized data, this data is deducted from the detail data. Thanks to this, the analyst doesn't need to analyze the detail data when he has to do a superficial analysis.
- Historical data, this is a collection of snapshots that the data warehouse makes at certain points in time. This way the DWH creates historical data and free up space for new detail data.
- Metadata, this data is a roadmap for data analysts. These are data about the data, this data makes it that an analyst can search for "client name" instead of including all the client names in his search command.

Now that we know which data can be found inside a DWH we are going to discuss how a DWH is composed. A DWH consists of 9 layers:

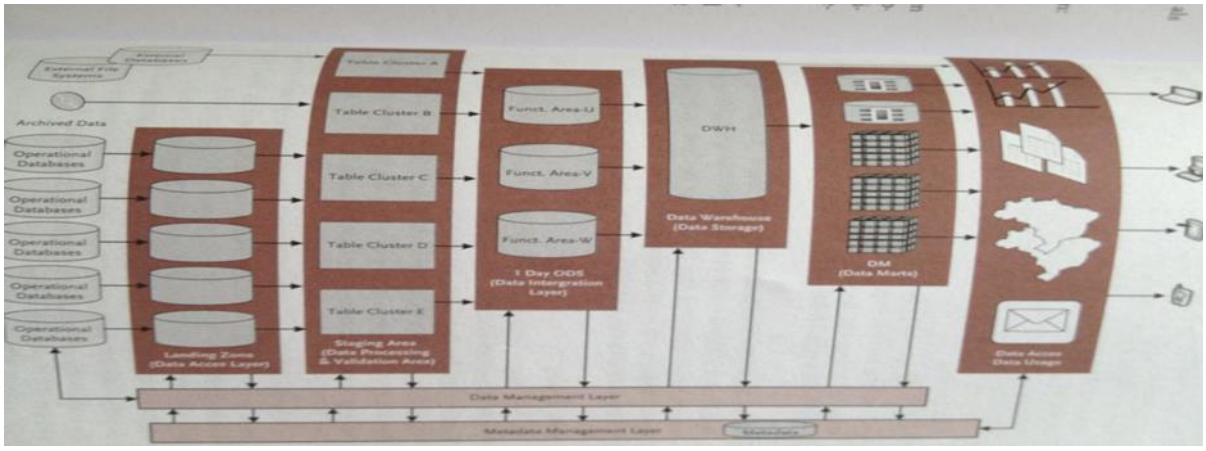

Figure 13 Layers of a data warehouse

- The data sources, the company gets his data from several sources. We can divide these sources into 3 groups:
  - Operational data, data that is generated by operational systems like MRP, ERP, CRM,…
  - Archived data, data stored onto USB-sticks, cd's, etc.
  - External sources, data coming from outside the company. This data is mostly bought by the company.
- Data extraction layer or Landing zone, in this layer we import and collect all the data from the different sources trough SQL commands.
- Data staging layer, all the data gets equalized through editing, filtration, etc. All the data that doesn't meet the norm will be deleted.
- Data storage layer, this is the core of the DWH. It is here that all the transformed data from the data staging layer is stored.
- Data marts, is a subset of the data storage layer. Data marts are used to quickly gain access to data from a certain subject. Thanks to these data marts the analyst doesn't always need to analyze the whole core.
- Data access layer, thanks to this layer the analyst can reach the data inside the DWH.
- Data Usage, is a collection of tools to analyze the DWH, there are 3 types of tools:
- Query and Reporting tools provide the option to make reports based on query command.

- OLAP, is used to collect data.
- Analytical Intelligence can find patterns in the data.
- Data management layer includes all the DWH management tasks. The data management layer has 4 tasks:
    - Give the command to the data extraction layer to import data.
    - To trace changes of data inside the DWH and correct them.
    - Back-up and recovery.
    - Manage query's to import data.
- Metadata management layer, the management of the meta data used inside the DWH.
- Transport layer; ensure the transport data between all the different layers.

6.1 How a data warehouse works

We will explain our self-made data warehouse in the chapter practical application.

But first we will explain how a data warehouse works in theory.

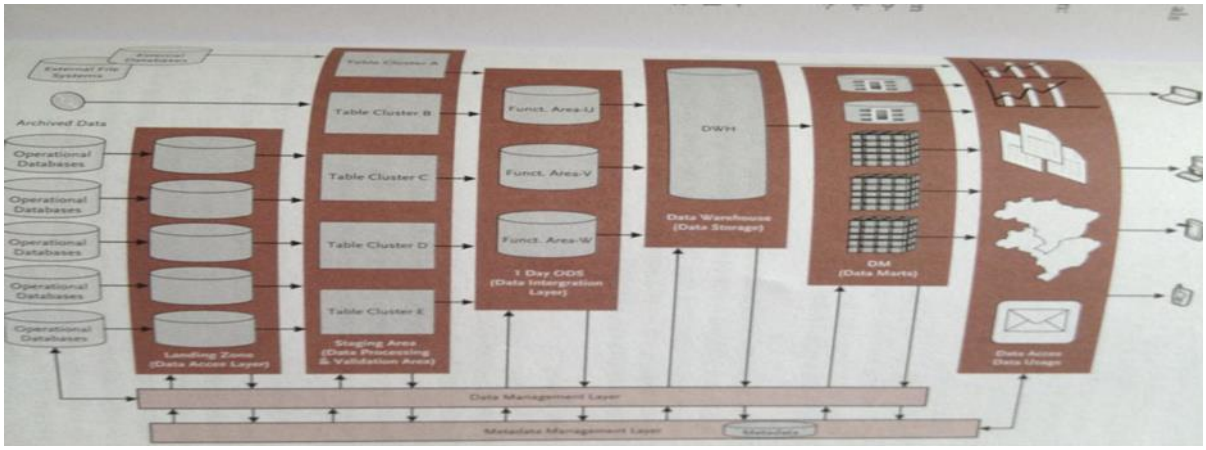

Figure 14 Layers of a data warehouse

The process on paper is relatively easy. We call it ETL.

ETL stands for:

- E – Extraction, get the data out of their sources/databases.
- T – Transformation, make the data useful.
- L – Loading, Store it into the database.

When explaining how the data warehouse works I added an E/T/L to the layers so you know in what “stage” of the ETL we are.

We have the operational systems in the company with each their own databases, we also have external sources with each their own external file databases and we have archived data. On a certain time planned by the data warehouse takes a copy of all this data and put it into the staging area (E). Thanks to the data access layer we can collect all this data through SQL commands (E).

Now that the data is in the staging area (T) all the data will be equalized and be in the same format and into one logical framework. All the data that isn’t up to the norm will be deleted. When this is all done, all this uniform data is stored into one large database thanks to the data integration layer (L) who made the storage possible.

But the database need have an answer for a set of questions, the marketing has their own set of questions, director of finance has his set of questions, etc. For each of them, we create a data mart. The data mart is a lot smaller than the database and delivers answers a little more quickly. When an analyst wants to analyze something in the whole database he can do this thanks to the data access layer.

6.2 The advantages of a data warehouse

6.2.1 Subject oriented

The DWH arranges their data around several subjects and these are stored into Data Marts. The advantage of this is that analysts can access the data they need even quicker. For example if an analyst wants to do an analysis on the profitability of their products he can just analyze the data marts that has this data instead of analyzing the whole DWH. (1keydata 2015)

6.2.2 Consistent data

All the data inside the company is processed and defined in the same way. This causes that all the data has the same value. For example the department account and financial planning are both working with euro instead of one working

with dollar and the other with pounds. The advantage of this is that all the data is somewhat equal and doesn't need that much transformation in the data staging layer. This saves a lot of time. (1keydata 2015)

6.2.3 Clean data

The management wants clean data from the DWH. With clean data we mean data that applies to certain rules. The advantage is that the "dirty" data becomes clean thanks to the DWH during the data staging process and if the "dirty" data doesn't want to become clean it gets deleted. This way the management always has clean data to make management decisions. (1keydata 2015)

6.2.4 Historical data

This is in our opinion the biggest advantage of a DWH. It gives the analyst the ability to analyze historical data. Thanks to this the analyst can find patterns or trends and use this to for example sell more products. (1keydata 2015)

6.2.5 A quick supply of data

A DWH is continuously supplied with the newest data to keep it as up-to-date as possible. This has as a result that the management always has the most recent data to analyze and make management decisions.

6.3 Disadvantages of a data warehouse

6.3.1 The data warehouse has possession over the data

The company loses control over the data. The DWH possesses all the data inside the company. This can cause security and/or privacy problems.

6.3.2 Expensive

The entry cost is very high. The IT infrastructure has to be reformed to support a DWH and the extra features implementation cost is very high.

7 DATA MARTS

In this chapter we discuss what data marts are exactly. This include which design should be used, the reason for creating data marts, the use of dependent data marts and the steps required for implementing, managing and /maintaining and creating data marts.

7.1 What are data marts

Data marts are the layer that lets users get fast access to data from a data warehouse so that they can use the desired data from the data warehouse. Data marts are connected to data warehouses and are mostly created for one purpose. That purpose can be for example for gathering data of a certain process of a company. A data mart contains only a little piece of the data that is stored in a data warehouse. (Swanhart 2010)

The different data marts that are used in an organization are often owned by the department that uses it. The ownership of a data mart includes not only the data but also the hardware and software used for that specific data mart. By giving the ownership of a data mart to a certain department, it allows the department to change, use, control and process their data to their convenience without modifying the data that is kept in the other data marts or in the data warehouse in general. This method of ownership is only possible on a data mart that is only used by one department. Examples of data marts that are often used by multiple departments are data marts for customers, sales and products. (Swanhart 2010)

Companies use data warehouses and data marts instead of simple databases because of the way that the data is saved in databases. Data in databases is not organized properly in comparison with data in data warehouses and data marts. This makes it very hard for organizations to search for the desired data that is kept in simple databases. Another problem is the time needed get the result of difficult queries. This is because databases are created to process a lot of queries and transactions on a daily basis what makes them slow. Another big difference is that transactional databases need to be updated while data marts

and data warehouse are designed to be read from. But there are also differences between data warehouse and data marts. Data warehouses were created to have the possibility to get access to giant amounts of records that are related to each other. (Swanhart 2010)

Using data marts allows reducing the user response time dramatically. Users get a fast access to the specific content of information they want to see. Data marts not only support the view of data by single users but also support the collective view of data by a certain amount of users depending on the hardware used.

We can see data marts as little data warehouses that contain fewer amounts of data and that are focused on a certain aspect of an organization. They reflect the supervision and process particularities of each of the organization's business units and are devoted to a particular function of the business. (Swanhart 2010)

It is possible to have various data marts containing data from different parts of a company to make them more accurate when linked together. If a company has multiple data marts, they can also use them together to better support the companies' business units. This allows business units to get more detailed and accurate information that is specific for them. An example of data marts that can be used together is the customer data mart and the sales data mart. When combined they give a precise view of which customer is linked to which sale. That kind of information is important for the sales or marketing department of an organization.

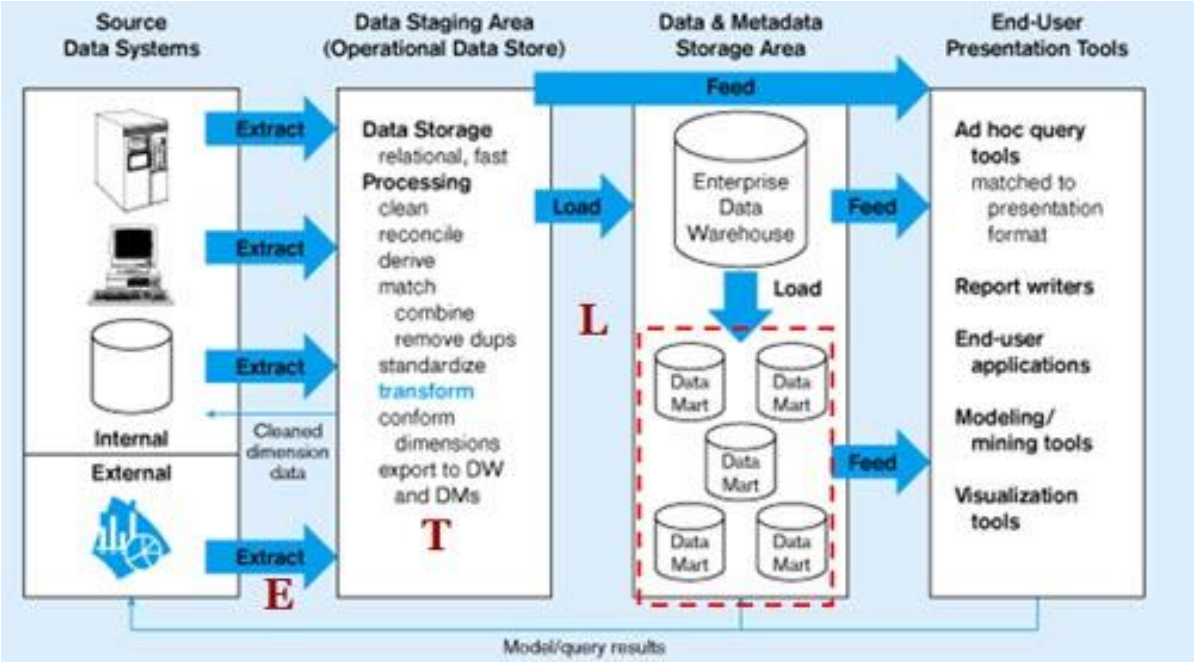

Figure 15 Data marts in a data system

The picture above (Figure 15) gives a good view of where data marts are placed in a data system. After extracting the data from the different sources, the data gets stored and processed to be able to be exported into the Data warehouse. The Processing of the data consists of multiple steps that change depending on the source of the data and the way it will be stored into the data warehouse. When exported, the data can be fed to the different tools (mostly ad hoc query tools) that the end users use. But to make it more efficient and fast, some parts of the data stored in the data warehouse that are often used can be placed into data marts. Then at their turn they feed the different tools like modeling/mining tools or visualization tools used by the end users.

7.2 Data marts vs Data warehouses

We now summarize the biggest differences between data marts and data warehouses. This comparison is done to give a better understanding of both data storing systems and why we should use one above the other in certain situations. For the comparison we compare both methods on: the number of subject areas it can hold, the data/information level, the data source integration and the model use.

Data marts:

- Mostly only hold one subject area (e.g. sales);
- The data hold may be more summarized than in a data warehouse but may also hold full detailed data;
- Focused on the integration of information given by one or a couple of data sources or of a certain subject area.
- Data marts are created around a dimensional model.
  (Standen 2008)

Data warehouses:

- Data warehouses can hold a multitude of subject areas;
- The information hold is very detailed;
- Is focused on the integration of all the different data sources;
- Isn't always created around a dimensional model. (Standen 2008)

7.3 When is a data mart needed

Some aspects of a data mart make it the best choice when data needs to be processed into useful information using different types of tools. Data marts make the accessibility to frequently demanded data very easy. This is because data marts are focused on a certain subject area. As mentioned before, data marts dramatically improve the response time of the user. Data marts are also easier to create than data warehouses. (Rainardi 2010)

The budget needed for implementing a data warehouse is for a lot of organizations way too high. Data marts are less expensive to implement and lets organizations to process their data into useful information that they can use to make decisions. (Rainardi 2010)

When the users need to be defined, then it is easier to clearly define the potential users using a data mart than using a data warehouse. Data warehouses contain a lot of different types of data and are highly cluttered. Where data marts only contain essential data and are not as cluttered as data warehouses.

Some other aspects can play a role when choosing between a data warehouse, data mart(s) or a combination of both but it is important to use the best system in regard of what we want to achieve with it. It would be a shame for an organization to put a lot of effort and money in the development of a full data warehouse while it doesn't really need it for the few amounts of data it has and the weak BI and predictive analytical knowledge of the users. (Rainardi 2010)

7.4 Possible Design schemas

Two kinds of schemas can be used when creating a data mart. The most popular is the star schema and the other choice is the snowflake schema. Both have advantages and disadvantages. However, the best choice and most used one remains the star schema because it allows a relational database to use analytical functionalities that are normally reserved to multidimensional databases.

7.5 Steps required for the implementing and creation of a data mart

For the implementation of a data mart we first need to create one. The creation and implementation of a data mart can be divided into five steps that have to be done in the right order. All steps are equally important because together they make sure that the created data mart will be working as attended and will keep working in the future. (Tsai 2007)

The five steps are:

- Designing the data mart;
- The construction of the storage;
- The use of sources to populate the data mart;
- Guarantee the accessibility to the data mart;
- The maintenance and managing of the system.

Each step will now be shortly discussed to give a better understanding of each of them.

7.5.1 Designing the data mart

The first step is designing the data mart. This step includes different sub steps that form the base of the data mart and are of great importance for the next steps. If some of these sub steps aren't done well, they will affect the whole data mart. As result the created data mart won't be working as desired and won't serve the right purpose. (Tsai 2007)

The sub steps are:

- Collecting the technical and business requirements;
- Defining the different data sources;
- Choosing the right data subset;
- Designing of the data mart's physical structure and logical structure;

7.5.2 The construction of the storage

With the construction of the storage we mean the creation of the physical database and also of the logical structure. These are created in association with the data mart. This way the data stored in the data mart will be easily, fast and efficiently accessible by the user (Tsai 2007). For achieving this, next tasks have to be done:

- Construction of the storage structure and of the database;
  This includes the creation of e.g. table spaces that are related to the data mart.
- Creation of the needed schema objects;
  This includes the creation of e.g. tables, indexes and links that were described during the 'designing the data mart' step.
- Decide which way is the most efficient to arrange the tables and the structure for accessing them.
  This step consists in choosing the right schema for the data mart like the star schema or the snowflake schema. Both will be described more in detail later in this thesis. This part depends highly on the different tables that are used and how they will be used by the end user.

7.5.3 The use of sources to populate the data mart

This step is needed to get the data from the different possible sources and to move it in the created data mart. This also includes the cleaning of the data, the reshaping of the data into the appropriate form and the adding of metadata. (Tsai 2007)

The different steps in the right order are:

- Defining the different data sources that will be used and targeting the structures;
- Data extracting from the selected data structures/sources;
- Cleaning the data from all impurities;
- Reshaping of the data in the appropriate form for the data mart;
- Loading the data mart with the transformed data;
- Creation of the needed metadata;
- Storing the created metadata on the right location in the data mart.

7.5.4 Guarantee the accessibility to the data mart

This step consists of using the data that is in the data mart to get useful information out of it. This is done by using queries on the data, by doing some analyzes on the data, by using business intelligence tools to create reports, graphics and charts. This step also includes the spreading of the information to the right persons so that they can make the right decision on the right moment using the information gained out of the data. (Tsai 2007)

If the end user doesn't use business intelligence or predictive analytics tools then it is needed to create a tool that will allow the end user to send queries to the data mart and to have a view of the accomplished result by the queries.

The tool is then a graphical front end tool that is the external layer for accessing the data mart using queries.

To enable the end user to access the data mart, a layer needs to be added between the external layer (the tool) and the data mart (the data). This layer is called the meta-layer. The function of the layer is to convert the data mart structure and its data into a humanly understandable form. This way it is possible for

the user to communicate with the data mart. This by using a language that is in accord and that is specific with the end user's function in the company. (Tsai 2007)

7.5.5 The maintenance and managing of the system

This is the last step after creating and implementing the data mart. The previous four steps combined the creation and the implementation of a data mart. The following step is the maintenance and managing of the system. This step needs to be repeated during the whole lifespan of the system to ensure its reliability, stability and necessity for the users. In short this step is about how the data mart needs to be managed during its lifecycle. (Tsai 2007)

Following tasks are part of the maintenance and managing step of a data mart:

- Managing and maintaining a secure data access;
- Control of the data growth in time;
- Constant amelioration of the whole system for higher performance;
- Guarantee of data accessibility during system errors;
- Redundancy of the data (back-up) for data retrieving.

8 SCHEMAS

In this chapter we go more in detail about the star schema and snowflake schema. Both are schemas that are used during the implementation of data marts. We discuss about the advantages, disadvantages, the different components, etc. of both types to give a better understanding of the use of them.

8.1 The star schema

The star schema is the easiest type of schemas for data marts. In short, star schemas are made of one fact table that are connected to one or multiple dimension tables. Star schemas are more efficient at handling simple types of queries than snowflake schema. We will discuss the other differences more in detail later. (Power 2008)

The name star schema comes from the fact that the shape of a star schema is similar with the form of a star as you can see in the picture (Figure 16). It is a star with a certain fact table in the center and multiple dimension tables around the center that correspond with the points of a star.

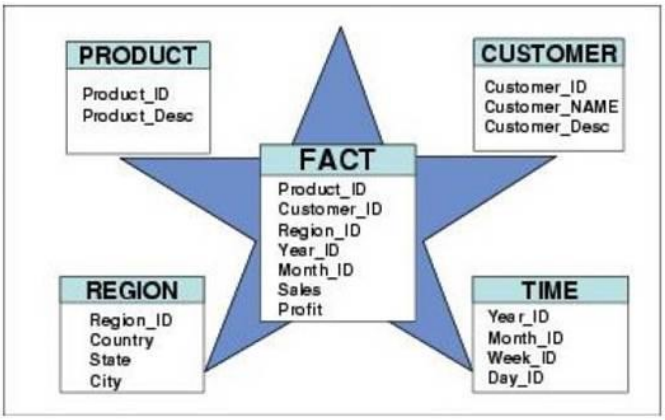

Figure 16 Start schema

8.1.1 Star schema model

The facts in star schemas represent the data of business processes. Those facts contain measurable and are linked to different dimensions. The measura-

ble are quantitative data that comes from the business process and the dimensions are attributes that are related to the fact data.

Examples of dimension attributes:

- Product colors;
- Product languages;
- Product types;
- Product sizes;

Examples of fact data:

- Time;
- Distance;
- Sales quantity;
- Sales price;

When a star schema has multiple dimensions, then it is gets the name of centipede schema. Star schemas with a little amount of attributes are easier to manage but are more difficult to use because the used queries needs to have a lot of joins to join all the needed tables for the query. (Wikipedia Community 2015)

8.1.2 What are fact tables

Measurements of a certain event are recorded into fact tables. Fact tables are created of foreign keys that are linked to dimensional data and numeric values. The dimensional data are where the information is hold. The design of fact tables is designed with a low level of granularity or gain. This allows the record of events at an atomic level by facts. As result of this atomic level record, the amount of records held in a fact table can increase rapidly over time. (Peterson 2010)

There’re three different types of fact tables:

- Snapshots: These are fact tables that keeps record of facts of a certain moment in time;
  An example is the production details at month end.

- Transactions: These are fact tables that keeps record of facts of a certain event;
  An example is the production events.
- Accumulating snapshots: These are fact tables that keep record of an accumulation of facts of a certain moment in time.
  An example is the total month-to-date production for a product.

To make sure that every row can be identified on a unique way, fact tables have surrogated keys assigned to them.

8.1.3 What are dimension tables

Compared with fact tables, dimension tables have a fewer quantity of records. The records of a dimension table have mostly a huge amount of attributes that are needed to define the fact data. Dimensions are used to describe a broad amount of different attributes (Peterson 2010). But dimension tables define mostly one of those kinds of attributes that are very common:

- Time: These dimension tables are used to define time at a very low level of grain for some events in the star schema that are recorded.
- Employee: These are dimension tables that are used to define workers/employees.
- Geography: This kind of dimension table is used to define a certain location like a city or a country.
- Product: These dimension tables are used to define a certain product.

Surrogated primary keys are mostly assigned to dimension tables. This in the form of a column of the type integer that is linked to the group of dimension attributes that combined creates the natural key.

8.1.4 Advantages of the star schema

Star schemas aren’t normalized instead they are denormalized. It means that all the rules that come with the normalization of transactional relational database aren’t as strict when designing and implementing a star schema. (Oracle 1999)

The main advantages of the star schema’s denormalization and the use of a star schema in general are:

- Queries are easier;
- Because the logic behind joins in star schemas is easier than the logic behind the joins of normalized transactional schemas, data can be found using queries that are simpler.
- Faster processing of queries;
- The performance of read only reporting tools can be much more optimized using star schemas instead of fully normalized schemas.
- Faster accumulation/aggregation;
- Because the used queries are simpler when using a star schema, it can ameliorate the accomplishment of aggregation processes.
- Cube feeding;
- OLAP systems use star schemas to create OLAP cubes more efficiently. Some OLAP systems uses some kind of ROLAP mode so that they don't need to build a cube and can simply use the star schema as their source
- Easier logic behind business report.
- Star schemas make the logic behind regular business reporting more simple compared to the logic behind the business reporting of fully normalized schemas.

8.1.5 Disadvantages of the star schema

Databases that are highly normalized have a strong data integrity enforced. This is not the case with star schemas where the data integrity is way less enforced. That causes it to be one of the main disadvantages of star schemas. Because of the strong enforced data integrity of normalized schemas, they are designed to resolve and avoid possible data anomalies that can happen after updates or wrong data inserts. Star schemas are because of their design more prominent to have data anomalies if the right precautions aren't taken. This is why the data inserted in star schemas are extremely well verified via batch processing, in real time or in near real time. This allows them to be protected against data anomalies even If they lack the protection that normalization procures. (Oracle 1999)

Another disadvantage of star schema is that they are not as flexible as normalized schemas when it comes on analytical purposes. Every possible analytical

query can be executed on normalized models. The only thing that the query needs to be in harmony with is the business logic of the normalized model. In comparison star schemas are created to serve a specific kind of data that wants to be viewed. This means that it is only possible to analyze the specific data that is hold in the star schema using specific queries. (Oracle 1999)

Star schemas are also not created to support some kind of relationships like the many-to-many relationship among different business units. These kinds of relationships are mostly transformed into a different form to be in harmony with the dimensional model of the star schema. (Oracle 1999)

8.2 Snowflake schema

Multiple tables of a multidimensional database that are placed in a logical arrangement so that the diagram of the relations between the different entities is in the shape of a snowflake are called snowflake schemas. Situated in the center of a snowflake schema is the fact table that is surrounded by the different dimensions with which it is linked as you can see in the picture (Figure 16).

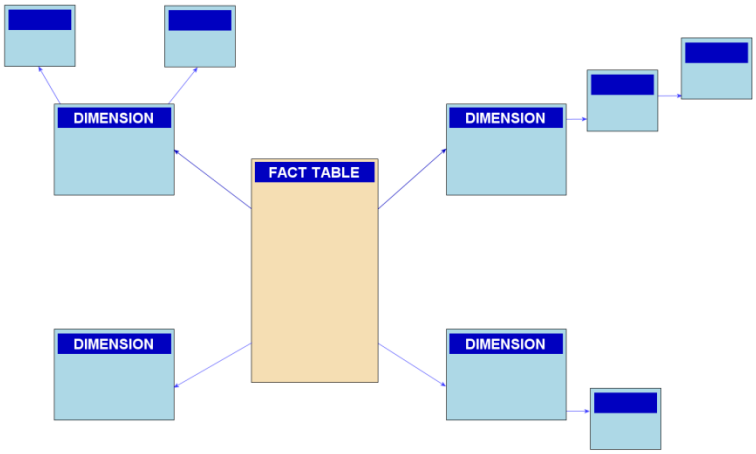

Figure 17 Snowflake schema

It is possible for a star schema's dimension tables to be normalized. When the star schema's dimension tables are normalized then the final configuration is in the shape of a snowflake with in the center the fact table. This method of normalizing the dimension tables of star schemas into snowflake schemas is called "snowflaking." The main idea when snowflaking is to get rid of the attributes that forms the low cardinality and to split tables of the star schema into new tables, this way the different dimension tables of the star schema are normalized.

A snowflake schema is in fact quite the same as a star schema. The difference is that the dimensions of a star schema are denormalized which means that one unique table defines every dimension, whereas in a snowflake schema because of the normalization of the dimensions, the dimensions are split into a multitude of tables that are related to each other. (Levene & Loizou 2011)

A snowflake schema can become very complex. This happens when the dimensions of it have a high amount of relationships with a multitude of different levels, when the dimensions are elaborated and that the child tables are linked to a lot of parent tables.

8.2.1 Use of snowflake schemas

As mentioned before the most common schemas for dimensional data marts and data warehouses are the star schema and the snowflake schema. These schemas are used when a low response time for retrieving data is of higher importance than how efficiently the data is manipulated. The data can be retrieved very fast because the schemas are not normalized at all (star schema) or are normalized but not more than the third normalization form (snowflake schemas).

When considering the deployment of a star or a snowflake schema, it is important to take in consideration the structure of the database platform and the tools that will be utilized for the queries.

Snowflake schemas are perfect for query tools that are highly sophisticated. Those tools mostly implement an abstraction layer between the raw table structures and the users. This allows the use of great amounts of queries with a high complexity. (Levene & Loizou 2011)

When the underlying table structure needs to be wildly exposed to the users, then it is recommended to use a star schema instead of a snowflake schema. Because star schemas can better handle query tools that expose the underlying of table structures and queries of a low complexity.

8.2.2 Normalization of data and storage of it

Normalization helps to avoid data duplication. This is done by splitting clusters of data that has the tendency to repeat often into different tables so that they are separated from the other data. Because of the normalization more tables are created. That implies that more joins have to be placed between the tables to be able to use certain kinds of queries. The advantage of normalization is that less storage space is needed to save the data (no redundancy) and it reduce the amount of locations that are needed to be modified when data has to be updated or changed (data is split).

Dimension tables are mostly way smaller than the fact tables. When snowflaking dimension tables the gain on saving storage space is often lost.

To somewhat resolve this problem it is possible to build an underlying layer that has the structure of a snowflake. Above this layer, views are placed that are used as the many joins that are needed to imitate a star schema. This way queries are as easy as when using a star schema and the benefits of normalizing dimensions is kept for the storage reduction. But this way of working is only possible if the used servers are powerful enough to automatically handle the underlying joins, that querying is simultaneously possible even if extra joins are placed between tables that are not needed for accomplishing some of the queries.

8.2.3 Advantages of snowflake schemas

Snowflake schemas and star schema are in fact created out of the same logical model. We can see star schemas as a distinguished condition of snowflake schemas. Snowflake schemas have some particular advantages that make them more suitable than star schema in certain situations.

The most important advantages are:

- Some kinds of OLAP modeling tools that are used in multidimensional databases are more efficient when using a snowflake schema;
- Because of the normalizing of the attributes a lot of storage space can be saved. This can be important if storage possibilities are limited.

8.2.4 Disadvantages of snowflake schemas

Because of the extra levels of attributes that are generated by the normalizations of the dimensions tables in snowflake schemas, the difficulty of the queries is higher when using joins to join different sources. The usage of snowflake schemas was some years ago very disapproved. The cause of it is that snowflake schemas are created to store normalized data in an efficient way without paying attention to the dramatic decrease in performance in the required joins for browsing dimensions. The critics about this problem have receded with the years because of the improvement of the query performance that are utilized in the browsing tools. (Wikipedia Community 2015)

Like star schemas are snowflake schemas not protected against data anomalies even if snowflakes are partially normalized. Only fully normalized transactional schemas have the data integrity security, which let them avoid data anomalies when inserting or updating data. Because of the lack of data integrity assurance, data that is inserted or updated in a snowflake schema needs to be manipulated and checked with precaution to avoid any data anomaly in the snowflake.

9 PREDICTIVE ANALYTICS

This chapter handles about Predictive analytics. We first start with explaining what predictive analytics is. After that we go deeper in the different techniques and usage of predictive analytics.

9.1 Explanation of predictive analytics

Under predictive analytics we understand the analyzing of data of now and from the past to create predictions about events that will happen in the near or late future. To analyze the data different methods are used like machine learning, statistical modeling techniques and data mining. (Redictive Analytics World 2013)

To be able to predict potential opportunities and risks, organizations and cities use predictive models. These models analyze patterns from data like transactional data and historical data. The models take different factors into account while capturing relationships. The relationships are then used to evaluate the risk or probability that is associated with a couple of conditions. (SAS 2015)

Predictive analytics is a part of the data mining area. Predictive analytics extract information from data systems to predict behavior patterns and trends. Predictive analytics is mostly used to predict an event that is in the future. But it can be also be used to predict unknown things from the past, present or near-future/future. (SAS 2015) Some examples are:

- Identifying credit card fraud while it happens (present);
- Identifying criminals after the crime was committed (past);
- Identifying new market possibilities for a product (future).

To explain it shortly, the fundamentals of predictive analytics consist of monitoring/capturing relationships of explanatory and predicted variables that occurred in the past. These captures are then used to predict outcomes that where unknown before the prediction. The results' accuracy depends highly on the quality of the assumptions and the grade of data analysis.

9.2 The different types

Predictive analytics is often used as term for forecasting, scoring data and predictive modeling. But it is also used to mention related analytical disciplines like optimization, decision and descriptive modeling. Organizations utilize these disciplines for decision making and segmentation by analyzing data very rigorously. The underlying techniques of the disciplines vary depending on the purpose. (Bertolucci 2013)

Some of the models will be explained now to give a better understanding of their purpose.

9.2.1 The predictive models

The purpose of the models of predictive models is to evaluate the chance that a certain unit that is placed in a different environment or sample will show the same specific performance as it did. In short, the models define the relationship between the specific performance of a specific unit in a certain environment and the environment. Certain of the units' attributes/features have to be known to be able to use them. The units are categorized into two groups. If the attributes and the performances of a unit are known then it is categorized as a "training sample". The other units are categorized as "out of sample" and are units that have unknown performances but known attributes. (Van Bochove 2014)

Predictive models have models for a wide variety of areas like:

- Marketing models: These models are used to find patterns in data that can provide answers about the preferences of customers. (SAP 2015)
- Fraud detection models: These kinds of models are used to find patterns in data that indicate/detect fraudulent behavior. (SAP 2015)

9.2.2 The decision models

Decision models define the interaction among all the components of a decision. The following components are necessary to predict the outcomes of decisions that have a multitude of variables:

- The decision;

- The forecast outcomes of the decision;
- The known information.

The use of decision models is mostly for optimizing or maximizing certain results while keeping other results identical or lower. These models are implemented by companies or cities to create decision logic or a list of rules that guarantee the achieving of the desired goal for every citizen, consumer or case. (Van Bochove 2014)

9.2.3 The descriptive models

The difference between predictive models and descriptive models is that predictive models concentrate on the prediction of the behavior of one single customer while descriptive models are focused on defining and quantifying relationships between different customers and products (data). That is why descriptive models are mostly used for the classification of customers or the prospecting of groups. An example of descriptive models application is the categorizing of the consumers by their life stage and product preferences. (Van Bochove 2014)

9.3 Applications of predictive analytics

Predictive analytics can be used in many different fields and applications. We will now discuss some examples of applications where predictive analytics is used and has shown positive effects.

The applications are:

- Analytical customer relationship management (CRM);
- Collection analytics;
- Cross-sell;
- Direct marketing;
- Fraud detection;
- Prediction of portfolio, product or economy.

9.3.1 Analytical customer relationship management (CRM)

Analytical CRM is frequently used in commercial applications. The techniques of predictive analysis are implemented on the companies' customer data to achieve customer relationship management's targets. The targets imply the creation of a holistic view of the companies' customers by using data from the company regardless where the information is kept or the department that is involved.

Predictive analysis is used by CRM for example in its marketing, services for the customer and sales. The predictive analytics tools help companies to optimize the use of their resources more effectively on their width range of customers/consumers.

Organizations have to know which products are demanded by their customers/consumers by analyzing them. Companies have then the possibility to predict the purchase habits of their customers/consumers to be able to sell and promote additional products when it is the right moment. They also have to recognize and reduce the problems that can lead to a loss of customers/consumers or that affect the companies' capacity to enrich their customer amount.

Analytical CRM can be used during the whole life-cycle of the customer by analyzing the changes in the relationship between the company and the customer, the purchases done by the customer, the retention, and the gains obtained by the company and some other important aspects. Some of the following applications that will be discussed are elements of the CRM.

9.3.2 Cross-sell

It is common for companies to store and conserve customer records, sale transactions and other abundant data. This data is then analyzed to find secret relationships that can allow a company to have an advantage over their competitors. When a company sells a multitude of products then predictive analytics can help the company to improve their cross sales or to additionally sell some other products to the customers. This is done by analyzing some aspects of the customer like his/her way of spending money, behavior, preferences and prod-

uct usage. By doing this the company can increase the profit it gains from each customer and it can enhance the relationship between the company and the customer.

9.3.3 Collection analytics

Portfolios mostly have a group of customers that never pay on time. The financial institutions are then obligated to have collection activities to recover the money that these customers should have paid. But some of the customers aren't willing to recover, this means that some of the collection resources are toss away without benefit. By using predictive analytics, financial institutions are able to better manage the use of collection resources, reduce collection costs and increase the recovery percentage. The predictive analytics tools will search for the most efficient collection agencies, legal actions and strategies in regard of the customer to optimize the change that the customer is willing to recover.

9.3.4 Direct marketing

Marketing is finding the best way to reach the target consumers for a certain product or service and to stay ahead of the consumers' behavior changes and market competitors. By using predictive analytics, companies can find what the best combination of timing, communication methods, marketing resources and products are to reach the desired consumer. Predictive analytics are mostly used to decrease the "cost-per-action" or the "cost-per-order".

9.3.5 Fraud detection

There are various types of fraud and fraud is a real plague for plenty of organizations. Some examples of fraud are: identity theft, online/offline fraudulent transactions and falsified credit cards. Not only big companies are affected by fraud but also small ones from different sectors. Examples of companies that are likely to be victim of fraud are:

- Insurance companies;
- Manufacturers;
- Services providers

- Retail merchants.

Using predictive analytics' models can reduce the chance that organizations are exposed to fraudulent activities. To recognize high-risk fraud possibilities in different sector or in general, predictive modeling tools can be used.

A risk-scoring technique has been developed by Mark Nigrini to recognize and examine targets. The technique will now be explained using an example of a fast food chain. Each restaurant is scored by a certain amount of predictors. The score of each predictor is then cumulated to give a global score to each restaurant. The total score of the restaurant represents the risk score. By comparing the different scores of the restaurants we can then predict which restaurant has the highest chance to be victim of fraud. This technique has also been used to identify the risk score of highly potential fraudulent travel agents, vendors and on divisional controllers to recognize fraudulent submitted reports

The evolution of the technology has recently made it possible to detect web fraud by using predictive behavior analysis. To predict web fraud the techniques use heuristics to be able to identify on the web, normal user behavior from abnormal user behavior that will certainly debouch into fraud.

9.3.6 Prediction of portfolio, product or economy

The customer is mostly not the main priority when prediction analyzes are done. The focus is mostly set on the firm, portfolio, economy or a product. As example we could take the Federal Reserve Board of a country. They could be curious in predicting the employment rate for the next couple of years or the prediction of the country's GDP evolution. Another example is the warehouse of a company. They would maybe want to predict the store level of products for their inventory management. This kind of desired information can be get using predictive analytical techniques like the regression technique or the machine learning technique. Both these techniques will later be shortly explained.

9.3.7 Risk management

The main purpose of utilizing techniques for risk management is to predict the future and to take the right decisions to get benefits out of it. The following ap-

proaches can be expanded from projects to market. They can also be expanded from near term to long term.

- CAP-M or Capital asset pricing model, this model predicts which portfolio will deliver the highest return;
- PRA or Probabilistic Risk Assessment, this model in combination with statistical knowledge predicts in a very accurate way risk forecasts;
- RiskAoA or Risk Analysis of Alternatives, this is a risk management tool developed by the United States Department of Defense. This tool allows the instant review of proposal, portfolio or alternatives risk and assess them to find the best choice.

The next subject that will be discussed is underwriting. Risk management is seen as a predictive method by underwriting and some business' approaches.

9.3.8 Underwriting

Businesses that are busy in some kind of sectors need to be able to predefine the cost that is needed to be charged to the client to cover a certain accident. We will now give some examples of companies that need that kind of information.

- Financial companies;
  Before giving a loan to a borrower, banks can use predictive analytics to have information about the potential ability of a borrower to pay the loan and an extra charge back in time.
- Car insurance providers;
  Car insurance providers can use predictive analytics to have information about drivers and cars so that they can set a specific price to cover each car and driver individually.
- Healthcare insurance providers;
  Healthcare insurance providers can use predictive analytics to predict how much it would cost a patient to have an operation in the future and charge them with right insurance price, this by analyzing data from their medical past.

By making predictions of health, crime, accidents etc. predictive analytics have the possibility to underwrite those here above mentioned quantities (price, risk percentage, success potential, etc.). By using predictive analytics companies are able to streamline their process for acquiring new customers. They can predict the level of risk that comes with a certain customer and how this risk will evolve in the future. The use of predictive analytics has changed the way banks approves loans to borrowers and how decisions are taken in the mortgage sector. It took sometimes days or weeks before pricing certain insurances or to give loans approvals. Now with the use of predictive analytics in the shape of credit scores, these things can be done in some hours. Using predictive analytics in a proper way can help to decrease potential future risks and can help to take the right decisions about service prices.

9.4 Predictive analytics techniques

The two main techniques of predictive analytics are the regression technique and the machine learning techniques. Both these techniques have sub-techniques that are all specialized for a certain type of application. The two main techniques will now be explained.

9.4.1 Regression techniques

Regression techniques/models form the main pillar of predictive analytics. The main purpose is to determine mathematical equations. These equations are then shaped into models that are used to define the synergy between the variables that are taken into consideration.

There are a lot of different techniques/models that can be implemented during predictive analytics. The choice of the model depends highly on the situation.

We will now give a short list of regression techniques/models to show the variety of them. The models/techniques won’t be explained because they’re not relevant for our thesis.

Examples of regression models:

- Linear regression models;

- Discrete choice models;
- Logistic regression;
- Time series models.

9.4.2 Machine learning techniques

Machine learning techniques are a part of the science that is focused on artificial intelligence. At the beginning, machine learning was used to create techniques, to give computers the ability to learn. Since a few years, machine learning techniques have found utility in a branch of fields like speech recognition, stock market, and fraud. This is due to advanced statistical methods that were added to machine learning for better regression and classification.

In some cases it is a direct prediction of the dependent variable enough, without the need of the different underlying interactions between the variables. But sometimes the underlying interactions are so complex and the dependencies unknown. In these cases, machine learning techniques will imitate the human way of thinking and will learn from exercising to predict the future.

Here is a list of machines learning methods that are often used for predictive analytics:

- Geospatial predictive modeling;
- Naïve Bayes;
- Neural networks;
- Multilayer Perceptron.

9.5 Predictive analytics tools

Predictive analytics tools have long been reserved to people who had the required competences and skills to use them and that were able to understand the results they produced. The new predictive analytics tools are now usable by almost everyone because they don’t require particular skills or competences.

Software companies that produce predictive analytical tools have tried to reduce the complexity of them by getting rid of the mathematical aspect and by adding a GUI that simplified the use for the user. They often also help the user to de-

cide which predictive model is the most appropriate for the loaded data. Users are then able to get meaningful information out of their data without analytical knowledge. The information is then displayed in a clear way using graphs, charts, reports, etc. to the user. For more advanced analytical experts there are still the more sophisticated predictive analytics tools. The higher sophisticated predictive analytics tools are often highly customizable and are able to process bigger amounts of data. (Wikipedia Community 2015)

Not only organizations but also cities are more and more using predictive analytics tools to support their decisions. This allows them to take decisions that will optimize the city and its surroundings on many levels like ecological, economical, technological and social level. They can do it now without outsourcing this to a company.

There are open source and commercial predictive analytic tools as well. Some open source tools are very well made and are of high sophisticated level. The communities behind them try to keep the tools updated and bug free. (Beal 2015)

Some examples of open source and commercial predictive tools are:

Open source:

- KNIME;
- OpenNN;
- Orange.

Commercial:

- BIRT Analytics;
- IBM SPSS Modeler;
- SAS Enterprise Miner;
- Dell Statistica.

IBM's, SAS' and DELL's predictive analytics tools package are the most popular and most used commercial predictive analytics software in the world. (Beal 2015)

10 SMART CITIES

Smart cities are something revolutionary that have already proved its effectiveness in different cities around the world. Governments and cities have realized that the data that they have generated and kept for so long can be used to make the city “smarter”. Thanks to all the technology we have described in our previous chapters, cities can reduce costs, enhance quality and optimize different aspects of the city. But how do these cities become smarter? Before we can explain it, we first need to explain the term ‘smart city’.

10.1 Smart City

A smart city is a city that uses digital technologies similar to the ones we have explained in our previous chapters (data warehouse, BI tools, data marts,…) to reduce the costs of certain tasks or jobs, reduce the consumption and improve the performance of urban services (For example optimizing the busses). We have to keep in mind that it isn’t just one BI solution implemented by the city that makes a city smart. It is the collection of BI solution in all the sectors and with transport and traffic management, health care, energy, water and waste as the main sectors. (Komninos Nicos 2013)

All these smart city solutions are developed to improve “the management of urban flows and allowing for real time responses to challenges”[18]. What this means is that thanks to all these solutions a smart city can act quicker to solve problems in the city and optimize certain flows to reduce costs and make the life of the citizens easier.

This all comes down to the definition of a smart city. Deakin defined a smart city in his paper as follow: “a smart city is a city that utilises ICT to meet the demands of the market and community involvement in the process is necessary

[18] Komninos, Nicos (2013-08-22). "What makes cities intelligent?". In Deakin, Mark. *Smart Cities: Governing, Modelling and Analysing the Transition*. Taylor and Francis. p. 77. ISBN 978-1135124144.

for a smart city”[19]. In other words a smart city wouldn’t be a smart city if it only had ICT technology. In order for a smart city to be smart I need to also implement the technology so it impacts the community. (Deakin Mark 2013)

10.2 How do cities become smart?

First of all the city has to implement IT into the city. This can be sensors, trackers, counters, etc. Before we can start with BI and analysing all the data we have to make sure that the city has a technological foundation.

When the city has the technological foundation the city can start to invest into BI solutions. This means implementing a data warehouse to collect all the data from all the technology or outsourcing various BI solutions.

This will enable the city to gain value out of their technological foundation, either with their partners or without them. Sometimes it is better for a city to outsource a BI solution since it is sometimes hard for a city to obtain certain funds to develop and implement a solution on its own.

When all these solutions are up and running the city is still not smart. A city only becomes smart when it uses all the reports and results from the various BI solutions that will be used to optimize processes inside the city and hopefully impacting the local community in a good way.

In our use cases we will explain how this work in real life and not on paper, we will also prove that these solutions actually reduce costs and improve the community in and environment in different aspects.

10.3 Characteristics related to smart cities

A smart city should use information technologies for different things. The use of information technologies in these different aspects of a city makes a city become a real smart city. (Gupta 2014)

---

[19] Deakin, Mark (2013-08-22). "From intelligent to smart cities". In Deakin, Mark.*Smart Cities: Governing, Modelling and Analysing the Transition*. Taylor and Francis. p. 15. ISBN 978-1135124144.

Information technology can be utilized for:

- The enhancement of the physical infrastructure of a city: This can be done using analytical methods and artificial intelligence to provide a city with a robust and stable development of its economic, social and cultural aspects.
- To enhance the partition of the citizens in the city's decision making and the governance in general: This can be done using e-participation, e-governance and processes focused on innovation.
- To enhance the learning, adaptation and innovation of a city to be able to respond in a more effective way to changes and circumstances. This can be achieved by optimizing the collective intelligence of the city.

The enhancement and optimizing of these aspects of a city result in the integration of three dimensions (the human intelligence, artificial intelligence and the collective intelligence) that are essential for a city. If we could see a city like a living being then the nerves would be the digital telecommunication networks, the (sensory) organs would be the different sensors and tags spread in the city and the brain would be the hardware and embedded intelligence. The cognitive competences and acquired knowledge would be the tools and software used for processing and analyzing the data received by the different parts. (Bhatt 2005)

Some forms of intelligence that smart cities uses, have already been showed. This is a short list of forms that have already been demonstrated:

- Empowerment intelligence: This type of intelligence has been shown in cities like Stockholm, Hong Kong and also in Melbourne. Empowerment intelligence consists of providing facilities, infrastructure and open platforms to regroup innovative ideas to specific locations in a city.
- Instrumentation intelligence: Instrumentation intelligence can be done by making infrastructures smart by collecting data, analyzing data and predicting things from the data received from these infrastructures in real time. An example of this type of intelligence is the use of smart electricity meters in homes to make citizens more aware of their real time consumption so that they can reduce it.

- Orchestration intelligence: The best example of this type of intelligence is Bletchely Park. This is the place where Alan Turing and his team have been working to decode the Nazi Enigma coder. Orchestration intelligence consists in establishing a place where a community can work together to resolve problems or to collaborate on things.

11 SMART CITIES -USE CASE

During our thesis we talked a lot about theoretical topics like big data, data warehouses, business intelligence, predictive analytics, etc. The question is how can we use all this knowledge and use it to gain an advantage or reduce costs/gain profits?

In order to show you that it is possible to use all this knowledge and how we are going to use it to gain an advantage or reduce costs/gain profits. We created use cases. Note that one of the use cases is still theoretical and do not describe the real situation of a city. The other use case is already in function in some cities around the world. We will describe later how these use cases can be implemented into the city of Joensuu.

11.1 Use-case 1: Garbage collection

The first USE CASE we developed is “Garbage collection.” We will first outline the current situation in most cities, after this is will go over what solution we have developed. When we have explained this, we will explain why we should implement this solution and lastly we will describe what the new situation looks like after implementing our solution and what the advantages will be. Please note that this is pure theoretical. The described situations are fictional but we tried to describe it as close as it would be in certain cities.

11.1.1 Current situation

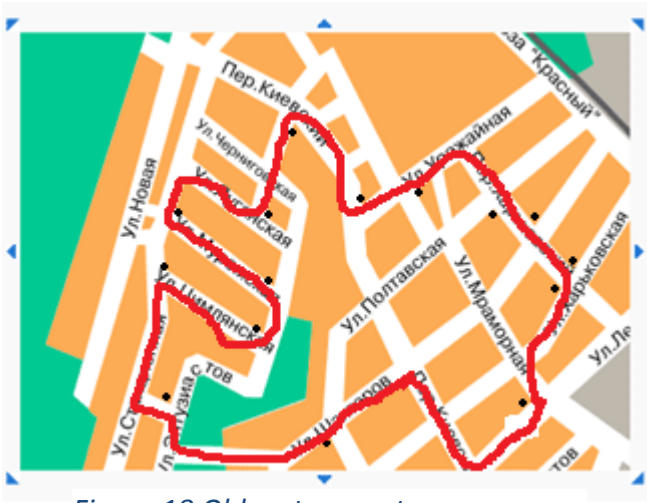

Figure 18 Old system route

A city has a lot of garbage bins. These garbage bins are placed all over the city. Every day, week, month, a city worker empties these bins. He follows a fixed route to empty all these bins. This means that every day, week, month he does the same route and it will take him approximately 2 hours to

complete this route. While doing this route, the worker empties 50 garbage bags and replace the 50 bins with new garbage bags.

As you can see in the picture, the black dots are the garbage bins and the red line is the fixed route he always follows. He follows this no matter what.

But can this not be more efficient and don't we lose valuable data when we don't store the data of how full the bins were? Before we will discuss this we would like to explain the solution we have created.

11.1.2 Solution

Our solution is pretty basic. We will implement wireless sensors into all the garbage bins in the city. The wireless sensors will send a value back to the city/waste collection company. The sensor will send the value every morning and every evening. This makes it so that if there was an event, even if the bin was empty in the evening but then full in the morning it will still be emptied.

This data will be stored into a data warehouse. We will use the data warehouse concept of our colleagues Jonas Lesy and Ruben Vervaeke. For the full explanation of all the components of the data warehouse and how to set it up we refer to their thesis "Applying Internet of Things". Now that we are able to store all this data we can start using it.

But before we can use this data we will first develop a route planner. This program is able to plan the most optimal route between several points and in our case garbage bins. The program uses "journey planner". A journey planner uses specialized electronic search engine. This enables the journey planner to find the best route/journey between two points. Before, this was mainly used by booking agents but now has found its way into the business world and now the business analytics world.

So now that we have the data warehouse and the route planner we can start working with our generated and stored bin values. Thanks to the reduce function that has been integrated in the data warehouse it is possible to search for bins that are 75% full or more. The reduce function will return all the bin numbers and how full they are. These bin numbers are now used by the route plan-

ner to calculate the most optimal and efficient route for the worker to collect these bins.

Over time we will have a nice collection of bin values and one of the functionalities of the data warehouse is to generate historic data. We can now use this historic data to analyze which bins are always full and look into the cause why the bin is always full and if we need to place extra bins in that area.

11.1.3 New situation

Now that we have explained our solution, how will this affect the old situation? Well based of the subtitle it will affect the old situation in a way that a new situation is created.

In the new situation every bin in the city has a wireless sensor. The sensors report back to the data warehouse how full the bin is. Based on this the route is made and the worker knows which bins he needs to empty and how he has to drive to complete his collection round.

Figure 19 New system route

As you can see in the picture, the green bins are bins that are 75% full or more. All the other bins are less than 75% full and aren't worth to be collected. This makes it that the route is much shorter than the previous one. The previous route took 2 hours to collect. With this route it will only take the worker one hour to collect all the bins and the good thing is that this route isn't fixed. So on a good day the worker only needs 30 minutes to collect all the bins. The work time will differ from route to route but our guess is that the average time will be 45 minutes to collect all the bins.

This is the main improvement in comparison with the old system. This enables the city/companies to save money on work hours and garbage bags. Next we

will calculate the benefits in the next subchapter the advantages of the new system.

Thanks to the data warehouse, the new system allows us to analyze the stored data for certain patterns. The way we use this in the new situation is to analyze all the historic data and look if we can find any patterns such as the bin at the marketplace has been full for the past 2 months. Therefore it might be a good option to put an extra bin there so the place will look cleaner because a full bin looks dirty.

11.1.4 The advantages of the new system.

The main advantages of this new system are reducing the costs of collecting the garbage. There isn't really a way to make profit with collecting garbage unless the city/company can sell the garbage which is highly unlikely. So this is why there are only costs reductions in our calculations.

The main cost reduction is the time the city/company saves when collecting the garbage bins. We know that in the old system it took the worker 2 hours to complete one collection round.
Thanks to the new system it takes the worker 30 minutes to 1 hour to empty all the garbage bins. In our calculations the worker does this round every day. This means that the city/company saves between one hour and 1.5hours in salaries. We are aware that this doesn't seem like a cost reduction since the worker probably gets paid for each day he works instead of the hours he worked. We will calculate the cost reduction as if the city/company would only pay the worker to do the collection of the garbage bins.

In this example the worker earns 10 euro/hour.

Old system	New system	
2 hours	1 hour	30 minutes
10x365x2= 5840	10x365= 2920	10x365/2= 1460

Table 1 Work hour cost old system vs new system

Our new system saves the city/company on average one month of salary. This save in salary isn't the only benefit that comes with this new system. A worker

probably works 8 hours a day, and 2 hours of that day he “wastes” on collecting garbage. With the new system we cut this waste in half so now the worker has 7 hours to do his other tasks. This increases the return of investment for the company and maybe earn them more money if he can work 7 hours on a profitable task instead of only 6 hours.

Another cost reduction that could be possible with our new system is reducing the amount of used garbage bin bags. Our new system cuts the costs for garbage bags thanks to our route planner. This makes it so the worker doesn’t have to replace empty garbage bins bags with new bags.

In this example there are 15 garbage bins in the old system and 7 in the new system.
A new garbage bin bag costs 4 euro.

Old system	New system
4x15= 60	4*7=28

Table 2 Bin bag cost old system vs new system

The new system saves on average 32 euro a day which makes a cost reduction of 11 680 euro a year. Now we do know that 4 euro for a garbage bag is probably too much but if you look at the whole picture this is definitely a big cost reduction for a city that has over 100 garbage bins.

These 2 cost reductions are the main reasons why every city should implement our solution if they haven’t already.

Another reason why every city should implement this is the environment. At the moment most cities feel dirty because they have a lot of garbage in certain places and full garbage bins. With our system it is impossible for the garbage bins to be full all the time and if they would be full all the time we would notice it when analyzing the data and recommend the city to place more garbage bins in that area. We can also analyze if certain garbage bins are full after certain events so we can find some pattern, a garbage trail, and based on that place temporary bins to collect all the garbage.

These are all benefits but everything has his downsides and so does our solution. Luckily there aren’t as much downsides as benefits.

11.1.5 The disadvantages of the new system

The main disadvantage is the cost. Cities and companies don't like to invest in IT unless they see immediate success or profit. Well our solution doesn't have that. The only thing is does is reducing cost and improving the city's image.

Our new theoretical solution will cost the city quite some money. The amount fluctuates because of the price of the sensors we have to implement.

Basically our solution has 2 components:

- Data warehouse
- Sensors

The cost of the data warehouse is going to be around 30 000 euro. This covers the setup and the implementation of the data warehouse into the system of the city/company. This is a rough estimation since we don't have to knowledge what servers will be needed if we implement this in a big city with a lot of garbage bins. We asked our colleagues Jonas Lesy and Ruben Vervaeke if this estimation was close to the cost they would have to make if they wanted to implement their own data warehouse. They both agreed that 30 000 euro is a good starting point.

The other costs are the sensors we have to implement in all the garbage bins. This will be the cost that determines how high the total cost will be. We think that the cost of a sensor is going to be around 50 euro. This can be reduced if we are able to mass produce these sensors. But if we use price 50 euro the sensors all together will cost 15 x 50 = 750.

So the total cost of our solution will be around 30 750. This may seem like a lot of money with no profit in return but is this actually true? According to the calculations in the advantages the new system pays itself back in 10 years (3000 euro less costs * 10 = 30 000). This may seem like a lot but you can't forget that the new system will also improve the environmental image of the city which could cause an increase in tourists and we all know how important tourists are for a big city.

11.2 Use case 2: Predictive policing

The second use case, predictive policing, is one that has been implemented in Amsterdam and in other cities. We will explain how the use case has been implemented in the city of Amsterdam. The use case is divided into different topics to give a better understanding of the whole solution. We explain what predictive policing is, what the problem was, how the problem has been resolved, and what are the advantages and disadvantages of the solution. The use case is closed with a conclusion that includes our opinion about it.

11.2.1 Predictive policing

For the matter of this case we first explain what predictive policing is. Predictive policing consists of utilizing predictive analytics and its different techniques to recognize and predict possible activities that are crime related and so on reinforce the law. Predictive policing's techniques are split into four groups. (Walter et al. 2013)

The techniques are split into following groups:

- Techniques for predicting crimes;
- Techniques for predicting offenders;
- Techniques for predicting perpetrators' identities;
- Techniques for predicting victims of crime.

Predictive policing is something revolutionary that has already shown its efficiency in some cities. But it is still not like a crystal ball that can foretell with certainty the crimes that will happen in the future. (Walter et al. 2013)

11.2.2 Amsterdam – Smart City

Predictive policing has been implemented in Amsterdam to help against the raising crime rate. The implementation is also part of Amsterdam's smart city project. The project consists of different implementation that affects some parts of the city and its citizens. Amsterdam is now highly occupied with following implementations: (Amsterdam City 2015 – Cohen 2014))

- Smart Mobility: Providing a safe, comfortable and efficient transport structure that is linked to an ICT infrastructure and open data;
- Smart Living: Providing a pleasant city to live, work and life by improving the quality of life including health, culture, safety and tourist attraction;
- Smart Society: A smart city is nothing without a smart society, that is why it is important to provide places for social interactions between people and creativity;
- Smart Areas: Developing areas to create a sustainable and efficient way of using resources;
- Smart Economy: Amsterdam wants to have a smart economy by enhancing the attractiveness and competitiveness of the city. This can be done by encouraging innovation, productivity, entrepreneurship and international attractiveness.

Figure 20 Visual representation of smart city project

The picture above (Figure 20) gives a visual representation of the places where Amsterdam is busy with implementing smart city tools and techniques.

11.2.3 Crime problem in Amsterdam

A lot of cities are nowadays confronted with a raise of criminal activities. Some examples of crimes are domestic burglary, mugging, robbery and murder. Amsterdam is also affected by this problem but less than if they would not have implemented a solution.

The main problem was that the police weren't always able to intervene in time because they were too far from the place of the crime or because they were already busy with another incident. The police department started to work with a

priority system that categorized the crimes by gravity, which helped, but still couldn't diminish the amount of committed crimes. There was also a problem with the police presence that was too low in some areas, which resulted in an increase of crimes.

A solution for this problem was to hire more police officers, but this was not economically possible for Amsterdam and wouldn't have helped to lessen the numbers of crimes. So Amsterdam turned to technology in attempt to resolve the problem and to help the police department to work more efficiently.

11.2.4 Solution

In short, the solution consist in using predictive analytics and data mining tools to intelligently and efficiently allocate manpower where and when it is needed in attempt to lower the amount of crimes. (Police Department of Amsterdan 2014)

To make this possible the police department of Amsterdam has invested into:

- Data miners;
- Data analysts;
- Needed data mining software;
- Needed dedicated servers;
- Needed camera's and installation of it;
- Needed sensors and installation of it;
- Tutoring of manpower;

We couldn't find the exact invested amount that was needed for the implementation in Amsterdam, but by comparing the amount that other cities have invested in their predictive policing we could calculate that Amsterdam has roughly invested 70 000 euro for the software and a yearly 40 000 euro for the maintenance of the system. This still not include the hardware, tutoring and hiring part. The total cost could possibly exceed 1 000 000 euro what is common for that kind of implementation.

We will now discuss more in detail how the police department of Amsterdam uses its predictive policing system, or as they call it, the "Crime Anticipation

System.” But first we are going to explain how things happened before. (Police Department of Amsterdan 2014)

The planning of the manpower before the implementation of predictive policing was very random since they used ad hoc analyses and their ‘gut feeling’. The police analysts created maps with hot spots; an example can be seen in the picture (Figure 21). The hot spots represented the locations where crimes were committed the past months. Then, the police went nearby or in the middle of those hot spots to do their patrol. This wasn’t effective at all because it couldn’t prevent criminal activities as desired. So they made some changes. (Police Department of Amsterdan 2014)

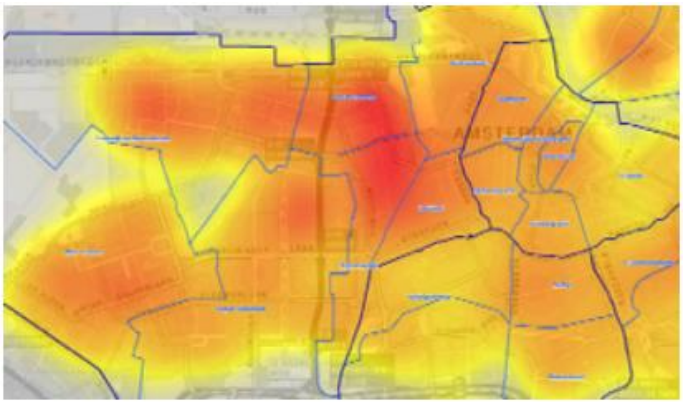

Figure 21 Hot spot map

The predictive policing changed their way of working for a more efficient one. The data mining/analyst team is now fully active on:

- The predictive policing;
- The uncovering of criminal networks;
- The extraction of useful information that comes from police reports and other data sources;
- The resolving of issues that comes from big data.

The first step, mapping:

Amsterdam has been divided into squares of 125m by 125m to create a giant grid that can be placed on the map of Amsterdam as shown in the picture (Figure 22). Some of the squares have been deleted because the majority of the square was filled with empty space (water, forest, pastures, etc.). After deleting these squares, the grid was reshaped containing 11 500 squares of 196m by

196m. This left them with an area of 38 416 m2 that needed to be invigilated, patrolled and controlled. (Police Department of Amsterdan 2014)

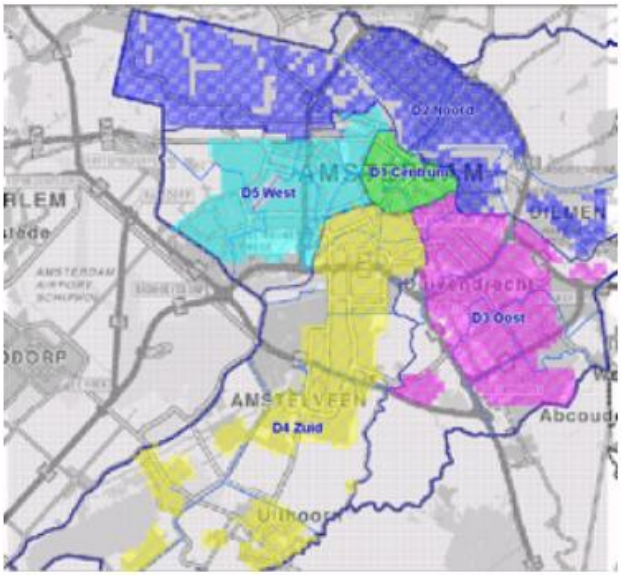

Figure 22 Mapping of Amsterdam

Second step, adding measures to the squares:

The squares were then filled with data that had been collected from the previous years that came from different sources like camera's, reports, and sensors, in which data was used as reference moments. Every square was composed of 78 data points that resulted in 897 000 data points in total. The squares kept information about the crimes that were committed in them (Crime history), the precise location of the crimes (location characteristics) and the crimes that happened within a number of weeks after the reference moment. (Police Department of Amsterdan 2014)

Third step, creation of an artificial neural network:

To link all the information together, an artificial neural network was created and implemented, this was done in cooperation with a software company like IBM that provide that kind of solutions. This resulted in a model that has the ability to give risk-scores to squares using information provided by data from the squares. (Police Department of Amsterdan 2014)

Fourth step, scoring the squares:

An example of scoring the squares for burglary:

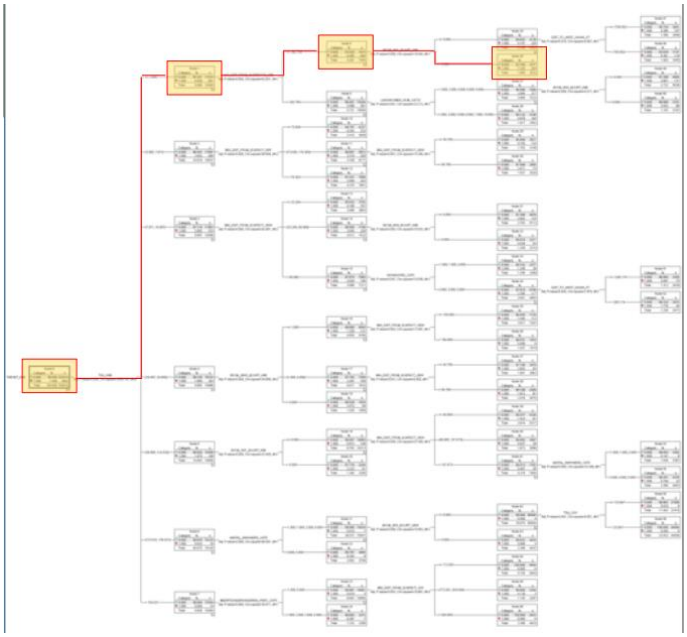

Figure 23 Scoring of map squares

Depending on the applied rules the scoring for burglary will differently change in the squares as you can see in the picture (Figure 23).

- If all the rules are rejected then the probability/percentage of burglary during the next x weeks will be the same in every square.
- If the last committed burglary in a square was less than a certain number of months ago then the probability/percentage of burglary in the squares that are in accord with the rule will rise.
- If in addition the distance to the closest x known burglars is less than x amount of distance then the percentage/probability of burglary in those squares rise also.
- If the last known burglary took place less than x weeks ago then the square's percentage/probability is raised.

The final percentage of the squares gives the probability/percentage of burglary in the square for the coming two weeks. (Police Department of Amsterdan 2014)

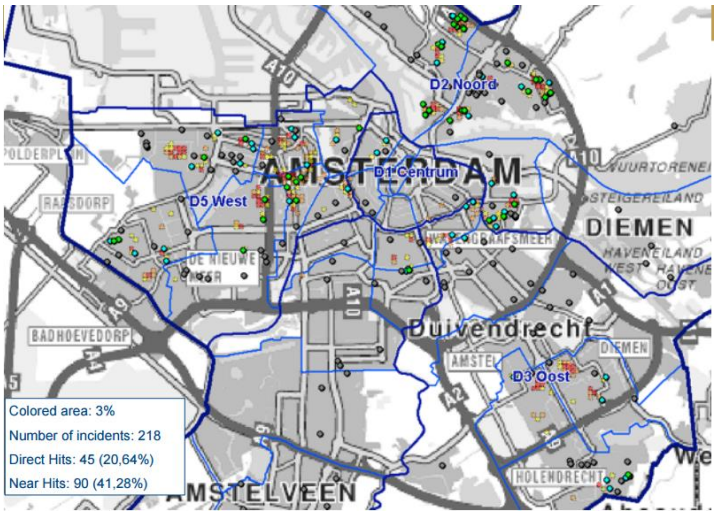

Figure 24 Visualisation of the data

This picture (Figure 24) gives a view of how the information is displayed to the police officers. By using this map, the police department can see where they have to patrol and pay attention. This is a prediction so it is still not completely reliable, though it does help the police department to do their job more efficiently and to better assign their manpower to strategic places to reduce the amount of crimes in squares and in Amsterdam in general. (Police Department of Amsterdan 2014)

Final step, automation and accessibility:

The whole system is process automated and doesn't need any kind of human interaction excluding system failures. The data is collected, transformed, and the needed models and maps are created every x weeks (2 weeks in Amsterdam). The maps and other information can then be accessed easily by the police officers using a HTML page that gives the user the requested maps. (Police Department of Amsterdan 2014)

11.2.5 Advantages

The use of predictive policing has had a great impact on the crime rate in Amsterdam. Amsterdam is still not crime free but if they wouldn't have invested and implemented their predictive policing, it would have been worse nowadays.

So the advantages are:

- Reduce of crime activities;
- Better and more efficient police patrolling;
- Easy access to the desired map information;
- Preventive crime fighting;
- Faster police intervening;
- Reduction of pressure on police officers;
- Use of artificial neural network;
- Automated system;
- Less police officers needed;
- Economical advantageous for Amsterdam in a couple of years;
- Example for other cities;
- Achieving their goal to become a smart city.

Since the implementation, the number of crimes has decreased which ranked Amsterdam at number 11 of the most livable cities in the world in the annual Quality of Living survey of 2015. That is already one place higher than before. (Mercer Company 2015)

Numbers of crimes in Amsterdam during 2013:

- Mugging – 2358 times;
- Robbery – 276 times;
- Burglary – 8257 times.

The numbers are lower than the previous years. But there's still a lot of work to do before eradicating criminal activities. (Police Department of Amsterdan 2014)

These are some of the advantages that came with using their predictive policing. The implementation is still very basic but in the future it will be more enhanced with new technologies and techniques that will fight crime before it maybe even happened.

Other cities have also implemented a predictive policing system and these are the experienced results. The Los Angeles Police department has seen its effectiveness and accuracy to be doubled compared to their old system. The police

department of Santa Cruz in California had a drop of 19 percent in the amount of burglaries during a test period of 6 months. In Kent, the street crimes occurred mostly in locations that were predicted by their predictive policing system, the successful percentage was higher than the one of their police analysts. (Wikipedia Community 2015)

11.2.6 Disadvantages

There are also some disadvantages related to this implementation. Amsterdam has certainly invested a lot of money into the implementation. Even if we don't have the exact numbers, when we see how much other cities have invested for their predictive policing we can imagine that Amsterdam invested almost the same or even more. Predictive policing is still something new that needs to be perfected. It is still not able to predict with hundred percent certainties where and when a crime will be committed and by who. At the moment predictive policing guess where and when crimes will be committed using mathematical techniques and different methods. (IBM 2013)

In a time where it is hard to find a job, the use of predictive policing reduce the amount of needed police officers. This is an advantage for the city because it reduces its police department costs. But this is also a big disadvantage because it increases the unemployment rate of the city what reduce the economical level of the city and make the city expend more money into their social department and tutoring activities.

The numbers of crimes has decreased in Amsterdam city center thanks to the system but the global amount of crimes has stayed quite stable. This is due to the increase of crimes around the city and in the country side. This is a disadvantage for the people living there. (Police Department of Amsterdan 2014)

Another problem is the camera and sensor surveillance. For some people the use of cameras and sensor goes against the right of privacy. This right is one of the most important one of our society because it is in harmony with the right of being treated as a human in any circumstances. If in the near future people are arrested even before they committed a crime. How can they possibly be sure that the person would have committed the crime? These are all problems and

disadvantages that will come with the future enhancement of this technology. (IBM 2013)

11.2.7 Opinion

We can conclude that the implementation of the system has its advantages and disadvantage. Some other cities and Amsterdam, in general, experienced positive results while using the system. We think that system still needs to be perfected and it also needs to be implemented in every place of a country to be sure that the crime rate is reduced and not moving to a place where the system is not implemented.

The cost of the implementation is quite high but it will become a must for every city and country to have that kind of system. There will be also some ethical questions/problems that will have to be resolved in the future to make sure that the system doesn't infringement some laws.

12 CITY OF JOENSUU – JOB REQUIRED

Now that we have theoretically explained our use cases, we applied these use cases to the City of Joensuu. Those two use cases are Garbage collection and Predictive policing. We start with the Garbage collection and end with the use case Predictive policing.

Note that the garbage collection will use the data warehouse that is developed by our colleagues Jonas Lesy and Ruben Vervaeke. The Predictive Policing use case will also use the data warehouse and we will explain step by step how to implement it into the city of Joensuu. We also provide information about the needed companies and jobs to realize the implementation.

12.1 Implementation of the use cases

12.1.1 Implementation of the Garbage collection solution

At the moment, the company Puhas Ltd is responsible for the collection of the garbage in the city. The company offers a lot of garbage collection solutions and is paid by the city to collect the garbage out of the bins. After checking the bins we know that they don’t use a system similar to ours since there are no sensors in the garbage bins.

There are 142 bins in the city center of Joensuu. Around the market place itself there are 20 bins. Outside the city there are big garbage containers near the apartments and supermarkets to collect all the garbage. Based on this we would say that there are enough bins in Joensuu to make Joensuu look clean. From our own experience this is in fact true. So our system will probably not improve the environmental look since it already looks very good. But it will of course cut the costs from collecting all the garbage bins.

At the moment Puhas ltd makes sure that every week all the garbage in Joensuu is collected. This means the bins and the big garbage depots are collected once every week. We don’t know how efficient their system is but since we see them every week at the same hour collection the garbage out of the big containers, we guess it is a fixed schedule. So this means that collecting the

garbage bins is also a fixed route and schedule. With our system this can be better and more efficient.

The way we would implement our system is by starting to add sensors to every bin big or small. These sensors would send how full the bin is and this data will be stored into the data warehouse we would have developed. The data warehouse we developed can be found in the Puhas ltd Company and will be similar to the one that our colleagues Jonas Lesy and Ruben Vervaeke have developed for their thesis.

How much will this reduce the costs for the company? We don't know the specific details of the people that work for Puhas ltd to collect the garbage but we do know that there are 2 employees collect the big containers and two employees that collects all the small bins in the city center and all in all it takes 8 hours to collect all the garbage spread over a week. This would mean that our system could reduce the payment costs for the Puhas by a considerable amount.

Our solution will cut the time needed to complete a garbage collection round in half. This means that it will take the 4 employees 4 hours instead of 8 hours to complete a round. This means that every employee will save 1 hour on average. One hour means a cost reduction of 10 euro a week multiplied by 4 is a cost reduction of 40 euro a week. On a year basis, this means that Puhas will save 2080 euro in salary payments alone. Puhas will also save on fuel since it takes the old system 4 hours to collect all the big containers which means that it will use around 15 liter each trip since a garbage truck uses 1 liter each 2 kilometer. Our new system will cut this in half thanks to the route planner of our solution. So our solution will save Puhas on average 7 liters diesel which translates in a cost reduction of 10 euro each week and 52 euro each year and let's not forget it will also improve the environment because there is now less pollution than before.

Now onto the cost of the project for Puhas. It will cost Puhas 31 420 euro to implement all this. This may seem a lot but in 10 years this whole system is paid back thanks to the cost reductions, and if they market it well, they probably can get some kind of fund from the government since they are improving the environment in a way. All in all this solution doesn't look like it improves a lot but in

the long run it will definitely make a difference. If “Puhas” wants to implement this solution they will have to hire an data analyst with knowledge of how a data warehouse works. Because they can just implement the data warehouse but they will still need someone to make everything work and get the right data out of the data warehouse. The best option for “Puhas” is to hire someone with the same background as ours. The employee that works with this solution needs to know the IT part of a data warehouse but also needs to know the business side of the solutions. This gets often forget but it is as equally important as the IT side.

12.1.2 Implementation of the Predictive policing solution

Finland is known for its low crime rate. This doesn’t mean they shouldn’t improve their anti-crime methods. Criminal activities are rising in a lot of cities and we suppose it won’t be different for the cities in Finland. That is why they have to implement this kind of solutions.

The current situation in Joensuu is probably the same as in Amsterdam before the implementation of the predictive policing. The police department works with preplanned patrol routes to spread their manpower on different locations where they think it is needed. Those patrol routes are created using some kind of hot spot maps of the crimes of the previous x months.

For the implementation of predictive policing in the city of Joensuu, the city will have to find a partner like IBM to help with the implementation, the future maintenance of the system, and to provide the city with the needed software, hardware and other resources. The police department will also have to hire some data mining and predictive analytics specialist to be able to use the created system by themself.

The first step consists in gathering all the data that the police department has about crimes that were committed in Joensuu. If the data isn’t digitalized yet, the digitalization will be the main priority. When the data is digitalized and gathered, it has to be transformed into a certain form that is needed to be able to place all the data in the data system that our colleagues have built or that is

provided by the partner that the city has found for the project. The form of the data is then defined by the data mining and predictive analytics specialists.

The next step is the mapping of the city. The city is divided into squares. Each square is then filled with information that is related to the square. The filling of the squares with information is to provide the system with reference moments. After that happened, an artificial neural network has to be created that link all the information together. The artificial network is then able to score the square for different kind of crimes.

The scoring happens the same way as explained in chapter 11 ‘SMART CITIES – USE CASE’ part 11.2 ‘Use case 2: Predictive policing’. This is done by using mathematical and analytical techniques. Enabling or disabling rules change the score given to the squares and so on change the predicted probability that a certain kind of crime happens.

When all this is done, the only part that remains is providing access to the desired maps and information. This could be easily done by using the web service that our colleagues have created or by one provided by the partner. A HTTP page linked to the web service could then be used, with the use of a password and login name the police officer can search for the desired map.

The whole implementation of the system won’t be cheap but helps the city to become a smart city and to provide the police department with a new and more efficient way to do their job and to fight against crime.

The advantages and disadvantages of the implementation will certainly be quite the same as described in chapter 11 ‘SMART CITIES – USE CASE’ part 11.2 ‘Use case 2: Predictive policing’.

13 PRACTICAL APPLICATION

The practical application of our thesis consisted in the implementing of data marts. This chapter contains our reasoning behind the implementation and the problems we encountered what resulted in an unfinished implementation of the data mart.

13.1 Data mart implementation

For our project we use the data warehouse of our colleagues Jonas Lesy and Ruben Vervaeke. They developed this data warehouse for their final project and explained every part of it in their thesis. What they didn't develop were the data marts. We have discussed the data marts in a previous chapter and since we proved the importance of the data marts we will implement them into the data warehouse. In this chapter we explain how we were going to implement the data marts

The practical implementation of the data marts was planned to be executed during the first two weeks of June, this was due to some problems our colleagues encountered which delayed the practical part of our project.

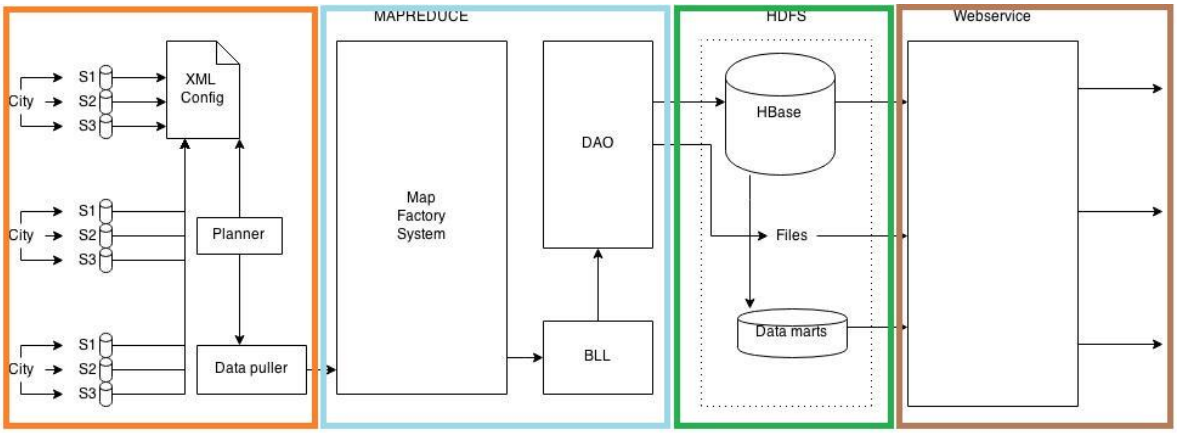

Figure 25 Concept of the data system

As you can see in the picture (Figure 25), we were going to implement the data marts in the HDFS part of the data system. The data marts would then receive their data from the HBase. The data that was going to be kept in the data marts

would then be accessible by Business Intelligence, predictive analytical tools and the web service of our colleagues.

The first step consisted in the analyzing of the data. The only data that we received from the city of Joensuu was data from some of their traffic lights that are shown in the picture (Figure 26). This meant that we could only implement one data mart. Even if a data mart was not really necessary at the moment because of the lack of data, we still tried to implement one to show the added value of using data marts in combination with a data warehouse.

Thanks to our colleagues, the raw data of the traffic lights that came from the city of Joensuu could already be saved in the HBase but without any processing of the data. They tried to work on the processing of the data in June but weren't able to finish it in time so that we could start with our part. The processing consisted to create the right tables, keys, joins, etc. in their HBase. At the moment of writing the data is just saved in the HBase just like in a simple database. Without this part done we were stuck because we weren't able to create a data mart that had the right fact table and dimension tables. The content of the fact and dimension tables needed to be in harmony with the different tables, key attributes, and column types of the ones in the HBase. So instead of discussing the real implementation we explain how we expected to do the implementation.

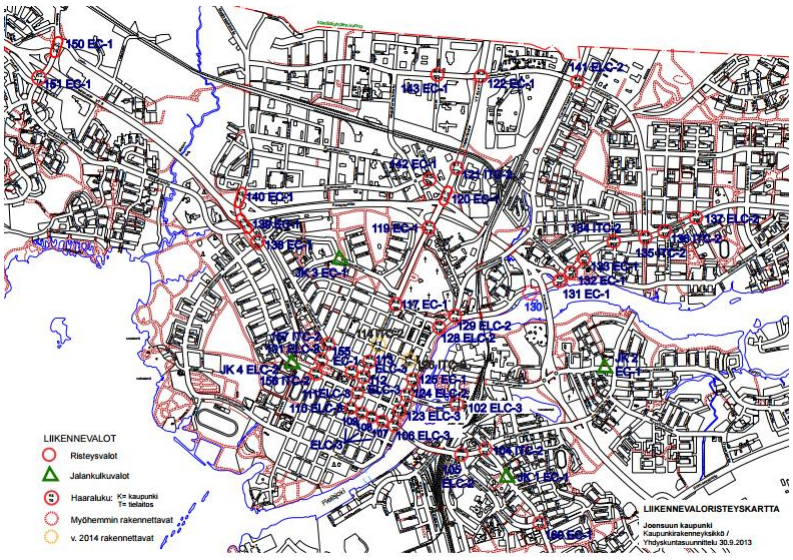

Figure 26 Traffic lights of Joensuu

The connection of the data mart with the web service was going to be the same as the one our colleagues were going to use for the connection of the HBase with their web service. So if Process Genius or the city of Joensuu wants to create a data mart they can easily use the same connection.

For the creation of the data mart we were going to use Hive mapper. Hive mapper is part of Hive. The graphical user interface of Hive is user friendly and easy to use as you can see in the picture (figure 27). That meant it could be flawlessly installed in the data system of our colleagues because it was compatible with Hadoop and Cloudera. After the installation we planned to create the right tables for the traffic lights data, this is where our project got stuck and couldn't be finished. One table would have been the fact table and woudl contain the fact data. The other created tables that would have been linked to the fact tables were the dimension tables. Both these types of tables were explained in the chapter about the start and snowflake schema. The creation of the different tables and the relationship between them should have been done using foreign keys and primary keys using Hive. The used keys would have been in harmony with the one used in the HBase to make the whole system very coherent. We were going to use Hive's data extractor to get the right data out of the HBase. But because the processing of the data wasn't finalized we couldn't do it. Once all the parts would have been created, the extraction of the data and the placing of the data in the right tables of the data marts would had happen automatically without the intervention of a human being.

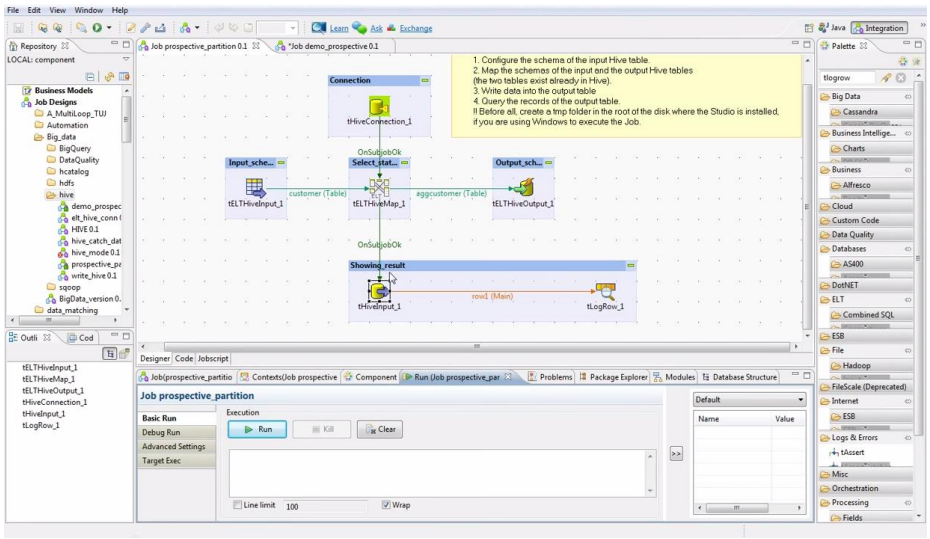

Figure 27 GUI of Hives

We know that this is not the real implementation of the data mart but we didn’t expect to encounter such a problem that would delay our whole practical part and wouldn’t let us finish the implementation of the data mart in time. Process Genius and the city of Joensuu can always contact us after our internship to help them to implement the data marts they need for some of their projects.

14 DISCUSSION/ CONCLUSION

Big data and BI solutions is definitely a value for companies and even cities, and it can be even stated that working with big data and BI solutions is a must. The benefits are not that significant for cities at the moment, but in the future, with even more big data, these solutions will be of great value. We hope we have proven this through our use cases.

When we were describing the big data we found out how “big” big data actually is and how valuable it is for the companies and cities. When researching big data we discovered that the business intelligence and the various BI tools use big data to create value data for the companies and cities. However, there was something missing. The big data was getting too big for traditional systems so we had to find an alternative way to store it. This was possible thanks to the data warehousing which is an interesting and future proof concept. Unfortunately, data warehouses also have their downside when it comes to BI tools that have to search through the whole database. This problem was easily resolved when we found out about data marts. We discussed data marts in a separate chapter because it is a very important part of a data system. After this, we talked about predictive analytics which is becoming very valuable in the business world and for cities.

After the background research and reporting, we had the necessary knowledge and examples to create or /explain some use cases where we actually could prove that big data, BI tools, predictive analytics, and a data warehouse can create valuable information for a company or a city. We think we accomplished this with our two use cases. Our tutor Petri Laitinen also asked us to implement the data marts into the data warehouse that our colleagues Jonas Lesy and Ruben Vervaeke had built. We described this task in in chapter 13 but were not able to implement it before the due date of our thesis

With our project we reached the following goals:

- Provided information about big data, BI, data warehouse, data marts and predictive analytics.

- Successfully created and /analyzed use cases based on the preceding research.
- Provided evidence through the use cases that big data, BI and predictive analytics can be beneficial for cities and companies.
- Gathered information on how we can implement data marts in a data system.
- Implemented data marts (Unfinished)

While writing our use cases we found out that it is much more complex than we initially thought. In addition to the IT limitations, we also had to take into account the business restrictions in terms of a budget, engagement and time to successfully implement a smart city solution.

We also didn’t expect we wouldn’t be able to finish the implementation of the data mart in time. The implementation should have been easy because we were well prepared for it. Unfortunately our colleagues couldn’t finish the total processing of the data in time before we ended our internship. We should have taken this into account and should have prepared our self with a back-up plan. But it was too late when we realized it. That is why time management is something very important when realizing a project in group.

15 REFERENCES

Gil press. (5/09/2013). *A very short history of big data.* Available: http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/2/. 13 may 2015.

Daan van Beek. (2014). *Wat is Big Data.* Available: https://www.biaward.nl/wat-is-business-intelligence-bi/. 13 may 2015.

Lisa Arthur. (15/08/2013). *What is big data.* Available: http://www.forbes.com/sites/lisaarthur/2013/08/15/what-is-big-data/. Last accessed 13 may 2015.

Meta group. (2001). *What is big data.* Available: http://www.neilstoolbox.com/bibliography-creator/reference-website.htm#. Last accessed 13 may 2015.

Thomas H. Davenport Jill Dyché. (2013). *Big data - Big companies.*Available: http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/bigdata-bigcompanies-106461.pdf. Last accessed 13 may 2015.

Wikipedia community. (2014). *Big data.* Available: http://en.wikipedia.org/wiki/Big_data. Last accessed 13 may 2015.

The elconomist. (2010). *Data, data everywhere.* Available: http://www.economist.com/node/15557443. Last accessed 13 may 2015.

Gary Simon. (2012). *Mastering Big Data: CFO Strategies to Transform Insight into Opportunity.* Available: http://www.fsn.co.uk/channel_bi_bpm_cpm/mastering_big_data_cfo_strategies_to_transform_insight_into_opportunity#.VWsyDc-qpBd. Last accessed 13 may 2015.

Laney Douglas. (2001). *3D Data management: Controlling Data Volume, Velocity and Variety.* Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf. Last accessed 13 may 2015.

John Webster. (2011). *Big Data: How New Analytic Systems will Impact Storage.* Available: http://www.evaluatorgroup.com/document/big-data-how-new-analytic-systems-will-impact-storage-2/. Last accessed 13 may 2015.

Kalil Tom. (2012). *Big Data is a Big Deal.* Available: https://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal. Last accessed 13 may 2015.

1keydata. (2015). *Data Warehouse Definition.* Available: http://www.1keydata.com/datawarehousing/data-warehouse-definition.html. Last accessed 15 may 2015.

Abramson, I (2013) *Data Warehouse: The choice of Inmon versus Kimball.* Available: http://ioug.itconvergence.com/pls/htmldb/DWBISIG.download_my_file?p_file=2346. Last accessed 15 may 2015

Jekel, R Simons, E Martin, A. (2014). *Using Proper Business Intelligence.* Available: https://www.atkearney.com/paper/-/asset_publisher/dVxv4Hz2h8bS/content/using-proper-business-intelligence/10192. Last accessed 15 may 2015.

Cardett Associates (2013). *Advanced Query Tool features* Available: http://www.querytool.com/features.html. Last accessed 15 may 2015.

SpreadSheet Concepts. (2013). *Advantages and Disadvantages.*Available: https://spreadsheetconcepts.wordpress.com/advantages-and-disadvantages/. Last accessed 15 may 2015.

Thomas C. Hammergren. (2014). *Querying and Reporting Tools for Data Warehousing.* Available: http://www.dummies.com/how-to/content/querying-and-reporting-tools-for-data-warehousing.html. Last accessed 15 may 2015.

Delfi : Report on the current status of the DELFI-system implementation. Federal Ministry of Transport, Germany and the Companies of DB AG, HaCon, HBT, IVU, mdv. Edited by Stephan Schnittger. 18 July 2006.” Last accessed 18 may 2015.

Wikipedia community. (2014). *Journey planner.* Available: http://en.wikipedia.org/wiki/Journey_planner. Last accessed 18 may 2015.

Chiang, A. (2011). *What is a dashboard.* Available: http://www.dashboardinsight.com/articles/digital-dashboards/fundamentals/what-is-a-dashboard.aspx. Last accessed 26 may 2015.

Wikipedia Community. (2015). *Data mart.* Available: http://en.wikipedia.org/wiki/Data_mart. Last accessed 8 May 2015.

Standen, J. (2008). *Data Warehouse vs Data Mart.* Available: http://www.datamartist.com/data-warehouse-vs-data-mart. Last accessed 8 May 2015

Swanhart, J. (2010). *Data mart or data warehouse?.* Available: https://www.percona.com/blog/2010/07/15/data-mart-or-data-warehouse/. Last accessed 12 May 2015.

Tsai, J. (2007). *Oracle Business Intelligence.* Available: http://docs.oracle.com/html/E10312_01/title.htm. Last accessed 13 May 2015.

Wikipedia Community. (2015). *Star schema.* Available: http://en.wikipedia.org/wiki/Star_schema. Last accessed 13 May 2015.

Power, D. (2008). *Data Warehouses, Schemas and Decision Support Basics.* Available: http://www.b-eye-network.com/view/8451. Last accessed 13 May 2015.

Utley, C. (2008). *Designing the Star Schema Database.* Available: http://ciobriefings.com/Publications/WhitePapers/DesigningtheStarSchemaDatabase/tabid/101/Default.aspx. Last accessed 13 May 2015.

Peterson, S. (2010). *Stars: A Pattern Language for Query Optimized Schema.* Available: http://c2.com/ppr/stars.html. Last accessed 16 May 2015.

Wikipedia Community. (2015). *Snowflake schema.* Available: http://en.wikipedia.org/wiki/Snowflake_schema. Last accessed 16 May 2015.

Levene, M & Loizou, G. (2011). *Why is the Snowflake Schema a Good Data Warehouse Design?.* Available: http://www.dcs.bbk.ac.uk/~mark/download/star.pdf. Last accessed 20 May 2015.

Oracle. (1999). *Oracle® Data Mart Builder.* Available: http://gkmc.utah.edu/ebis_class/2003s/Oracle/DMB26/A73318/schemas.htm. Last accessed 20 May 2015.

Wikipedia Community. (2015). *Predictive analytics.* Available: http://en.wikipedia.org/wiki/Predictive_analytics. Last accessed 20 May 2015.

redictive Analytics World. (2013). *Predictive Analytics Guide.*Available: http://www.predictiveanalyticsworld.com/predictive_analytics.php. Last accessed 20 May 2015.

Beal, V. (2015). *Predictive analytics.* Available: http://www.webopedia.com/TERM/P/predictive_analytics.html. Last accessed 21 May 2015.

Van Bochove, M. (2014). *Predictive analytics: kijken door de glazen databol.* Available: http://www.marketingfacts.nl/berichten/predictive-analytics-kijken-door-de-glazen-databol. Last accessed 22 May 2015.

IBM. (2013). *Predictive analytics.* Available: http://www-03.ibm.com/software/products/en/category/predictive-analytics. Last accessed 23 May 2015.

SAS. (2015). *What is predictive analytics?.* Available: http://www.sas.com/en_us/insights/analytics/predictive-analytics.html. Last accessed 27 May 2015.

Bertolucci, J. (2013). *Big Data Analytics: Descriptive Vs. Predictive Vs. Prescriptive.* Available: http://www.informationweek.com/big-data/big-data-analytics/big-data-analytics-descriptive-vs-predictive-vs-prescriptive/d/d-id/1113279. Last accessed 27 May 2015.

SAP. (2015). *Predictive Analytics Software.* Available: http://go.sap.com/solution/platform-technology/predictive-analytics.html. Last accessed 28 May 2015.

Karelia University of Applied Sciences. (2015). *Bachelor Degrees.*Available: http://www.karelia.fi/en/admission/studies/bachelor-degrees-in-karelia-uas. Last accessed 5 May 2015.

Joensuu City. (2015). *JOENSUU-INFO.* Available: http://www.joensuu.fi/. Last accessed 5 May 2015.

Process Genius. (2014). *About the Company.* Available: http://www.processgenius.fi/. Last accessed 5 May 2015.

Cohen, B. (2014). *The 10 Smartest Cities In Europe.* Available: http://www.fastcoexist.com/3024721/the-10-smartest-cities-in-europe. Last accessed 13 May 2015.

Gupta, M. (2014). *Smart jobs for smart cities.* Available: http://www.financialexpress.com/article/industry/jobs/smart-jobs-for-smart-cities/21796/. Last accessed 15 May 2015.

IBM. (2011). *Smarter Cities: Creating opportunities through Leadership and Innovation.* Available: https://www.ibm.com/smarterplanet/global/files/Budapest_MarchMF_12.pdf. Last accessed 25 May 2015.

Wikipedia Community. (2015). *Predictive policing.* Available: http://en.wikipedia.org/wiki/Predictive_policing. Last accessed 26 May 2015.

Police Department of Amsterdam. (2014). *CAS: Crime Anticipation System.* Available: http://event.cwi.nl/mtw2014/media/files/Willems,%20Dick%20-%20CAS%20Crime%20anticipation%20system%20_%20predicting%20policing%20in%20Amsterdam.pdf. Last accessed 26 May 2015.

Walter L. Perry, Brian McInnis, Carter C. Price, Susan C. Smith, John S. Hollywood. (2013).*Predictive Policing.* Available: http://www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf. Last accessed 29 May 2015.

Amsterdan City. (2015). *Amsterdam Smart City Project.* Available: http://amsterdamsmartcity.com/projects. Last accessed 29 May 2015.

Ganesh D. Bhatt (2005) Types of Information Technology Capabilities and Their Role in Competitive Advantage: An Empirical Study, Available: http://dl.acm.org/citation.cfm?id=1278006 . Last accessed: 4 June 2015.

Mercer Company (2015) 2015 City Rankings, Available: https://www.imercer.com/uploads/GM/qol2015/h5478qol2015/index.html. Last accessed: 4 June 2015.

IBM (2013) Predictive policing (IBM), Available: http://www.slideshare.net/socialmediadna/predictive-policing-ibm. Last accessed 25 May 2015.

Rainardi, V (2010) Reasons for Creating a Data Mart from a Data Warehouse, Available: https://dwbi1.wordpress.com/2010/03/03/reasons-for-creating-a-data-mart-from-a-data-warehouse/ . Last accessed: 15 May 2015.