Peng Gong

School of Electrical Engineering

Thesis supervisor:

Prof. Karlos Artto

Thesis instructor:

M.sc (Tech.) Belle Selene Xia

Aalto University

School of Electrical Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Peng Gong

Title: A Case Study of Data Analysis Process and Tools for a Consulting Company

Date: 23.05.2012 Language: English Number of pages: 73

Department of Communication Engineering
Professorship: Industrial Management

Supervisor: Prof. Karlos Artto
Instructor: M.sc (Tech.) Belle Selene Xia

It is crucial from any Consulting company's point of view to perceive some degree of data analysis in the environment of business intelligence. This research examines the different processes and tools in data analysis, and builds a specific and effective process and tool with cost-benefit analysis for Florilla Consulting, which can be beneficial to similar consulting companies operating in data analysis field.

Based on a large sample of qualitative data we demonstrate the benefits and importance of using data analysis processes and tools in business intelligence as a strategic necessity and show how this system can be implemented in various business case scenarios. Finally we propose a business model of data analysis to be tested by future research.

Keywords: Business Intelligence, Data Analysis, Data Warehouse, Data Mining, Data Mart, OLAP, Cost-Benefit Analysis

ACKNOWLEDGEMENTS

I would like to thank my supervisor Karlos Artto from Industry Management unit of Aalto University School of Science. Many discussions with you and your constructive comments have been very helpful during the thesis work.

I would also like to thank Florilla Consulting to give me this opportunity for my master thesis, and also my instructor Bele Selene Xia to enable and instruct this research.

Table of Contents

List of Figures

List of Tables

1 INTRODUCTION

1.1 Overview

This study aims to summarize the plan of conducting a research on different processes and tools in data analysis, and build a specific and effective process and tool with cost-benefit analysis for Florilla Consulting**,** which can be beneficial to other similar consulting companies operating in data analysis field.

Florilla Consulting is a consulting company with the knowledge of value network and the asset of data mining. Currently Florilla Product Family includes Data Forecasting and Trend Analysis, which all require an effective process and tool for data analysis. Meanwhile, data mining is also seen as an important tool by modern business to transform digital data into business intelligence, which gives an informational advantage. This especially benefits start-up companies like Florilla Consulting to compete against the big players in the competitive market within the same industry in terms of price, effectiveness and accuracy.

This research will help any consulting companies, especially Florilla Consulting to investigate more effective data analysis processes and tools, and find a specific process and tool via a cost-benefit analysis to conduct its business operation. The finalized process and tool will also enable Florilla Consulting to offer its clients a comprehensive understanding of the historical events and in formulating action plans for the future, while consolidating its core business. In addition**,** the finalized process and tool will as well meet the vision of Florilla Consulting, which is to offer the best solution to the clients at a reasonable price giving them an information advantage to their competitors and collective benefits of knowledge sharing.

The results of this research may be useful to the top management in Florilla Consulting, especially in their daily operations as well as their decision-making of business operations.

1.2 Research Objectives

The question of the research is *what would be a most effective data analysis approach from a tool and process point of view in Florilla Forecasting*? The objective is to do a research on different processes and tools in data analysis, then build a specific process and tool with cost-benefit analysis for a case in question. Furthermore, we will create a business model of data analysis for Florilla Consulting to test the business model with potential clients. The thesis work mainly focuses on the technical analysis from a business perspective, and the scope of the research is limited specifically to the case company.

While there are many researches done on how to produce successful data analysis processes and tools, we see that some aspects about the hectic information industry are more than an educated guess. This research is particularly relevant for the future success of Florilla Consulting and will benefit all small and middle-scale companies. In addition, the results of this research are mainly directed to the technology audience and tailored to the need of Florilla Consulting.

1.3 Data and Methods

The research is conducted with an inductive research approach. The literature used in this research originated from the well-known international literature on the topics of business intelligence, data analysis, and data mining with qualitative research. Moreover, various case studies in the past done on the software companies' act as foundation for the analysis. This study will be using the current data analysis methods and tools with a focus on evaluation plus the company-specific information that we gain from the representative also as this research instructor. We will also analyse the possible processes and tools that Florilla Consulting would use to tackle its daily operations via a cost-benefit analysis. This is important to see the different opportunities behind these processes and tools. At last but

not the least, we will build a suitable data analysis framework including process and tools for Florilla Consulting as a business model.

This research is of vital importance in revealing the effectiveness of the role played by data analysis in the future activities of Florilla Consulting.

2. BUSINESS INTELLIGENCE TECHNOLOGIES

2.1 Business Intelligence

The business environment has been constantly changing in decades, and it is becoming more and more complex. The well utilization of Business Intelligence has become the key for the business performance of any consulting companies and a necessity in the competitive markets. We think that understanding Business Intelligence and the advantage of Business Intelligence System would help Florilla Consulting conduct their business significantly. Therefore, we start the literature review on the topic of Business Intelligence first.

There has not been any unanimous definition of Business Intelligence, but we feel that the following two definitions describe the terminology well.

Evelson, Boris (2008) in Forrester Research gave the broad definition of Business Intelligence. “Business Intelligence (BI) is a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making.”

Lida Xu, et al. (2007) also mentioned the BI definition in a research for IEEE International Conference. “Business intelligence is the process of gathering enough of the right information in the right manner at the right time, and delivering the right results to the right people for decision-making purposes so that it can continue to yield real business benefits, or have a positive impact on business strategy, tactics, and operations in the enterprises.”

Business Intelligence is understood as a decision supportive system that ensures better business decision-making, and has an important role in the creation of information for operational and strategic business decision-making. We see that from Evelson’s definition of BI, the umbrella term business intelligence not only refers to an integrated set of tools supporting the transformation of data into useful information in order to support decision making, but also refers to the processes of data manipulation from preparation to

validation. Therefore, the common used terms like data mining, data warehousing, data mart, etc. all fit into the term “Business Intelligence” mentioned before. Furthermore, nowadays it seems that Business Intelligence is also concerned about organizational decision making processes, information and knowledge analytics and management, as well as human interaction. Hence, further related terms like, Business Performance Management (BPM), Decision Support System (DSS), and Management Support System (MSS) can all be classified under the umbrella term of Business Intelligence.

Ales Popovic (2010) mentioned a state-of-the-art Business Intelligence System in his research, a broader concept of BI system with a data-oriented approach, where the centre of the architecture represents integral data sources for analytical decision-making, and everything else like strategy, business process, and finance are all the elements of business intelligence. We think Figure 1 represents the broad concept of business intelligence well. The suppliers provide raw material or data as the source to the company, and the company running the daily operation is the centre in order to monitor the market trend and produce customized products for its customers with the whole business intelligence context.

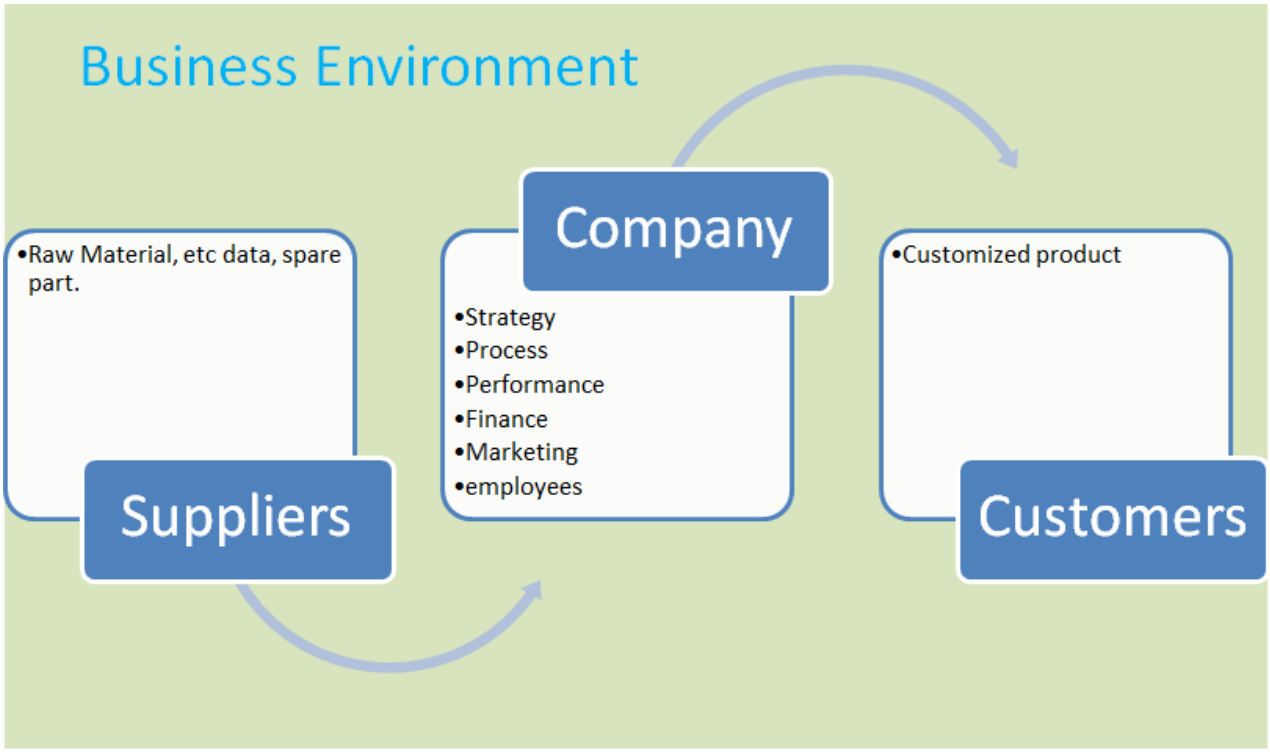

Figure 1: Business Intelligence Broad Concept (adapted from Ales Popovic, 2010, BI Broad Concept)

2.2 Advantages of Business Intelligence System

Business Intelligence as a system not only improves a company's business performance to achieve its objectives, but also has some other major advantages such as the follows:

- The implementation of business intelligence can make decision-making much easier.
- Managers can access and evaluate company's data at any given point of time and place for better decisions.
- Business Intelligence enhances managerial efficiency, enable employees to share data and simplify teamwork as well.

A research carried out by Thompson (2004) reported the following to be the major benefits of BI based on the results of a survey in terms of despondence.

Business Intelligence allows:

- Faster, more accurate reporting (81%)
- Improved decision making (78%)
- Improved customer service (56%)
- Increased revenue (49%)

Davenport (2010), Foley and Manon (2010) also stated that Business Intelligence applications have become the top spending priority of corporate information technology organizations, and Business Intelligence is one of the few areas of technology that are still growing**.**

Nowadays, most of the companies have their own BI models to handle daily-running business, and the Business Intelligence Models (BIMs) enable business people to transform enterprise data into business operation, strategies, and performance indicators through for example highly automated tools, Balanced Scorecard and Strategy Maps.

Turban, Sharda, and Delen (2011) proposed a state-of-the-art BI system including four major components in their studies:

- The data warehouse that accesses, processes the data source and returns the result. This is the heart of the BIS;
- BI analytical tools that manipulate analyse and mine data in the data warehouse, and return the integral view on the data to the users in case of for example reporting.
- Business Performance Management is used to monitor and analyse the business performance;
- A user interface is used to view the results returned from the data warehouse.

Based on Turban, Sharda, and Delen (2011)'s studies and general understanding of business intelligence, we draw Figure 2 to illustrates the overview of BI system architecture on a general level. The data coming from data source is extracted, transformed and loaded into a data warehouse, where data mining is manipulated. Then the result is demonstrated to the business user through user interface, as well as for the

executives and managers to monitor and analyse the business performance via BPM. This BI system is configured and maintained by different IT specialists who also give support to the business users.

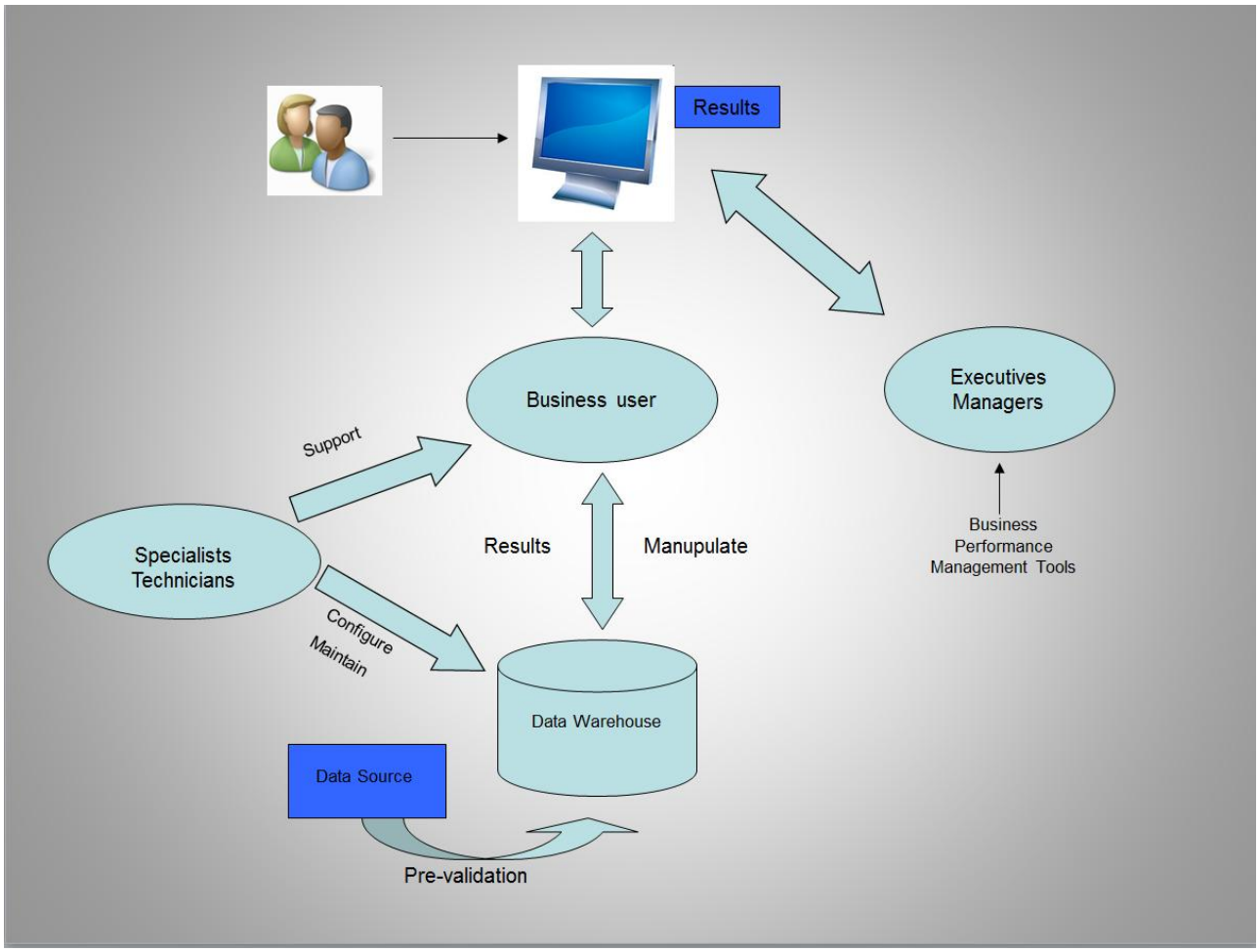

Figure 2: Overview of BI System Architecture

Any company that cannot manage the business with human support requires the implementation of computerized management system for the daily activities. BI allows the managers to make more informed and timely decisions based on computerized systems and tools, and the most common application areas of BI are sales and marketing analysis, planning and forecasting, finance, general reporting, and profitability analysis.

For example, the most widespread systems used by the various companies to handle finance / accounting and marking / after-sales nowadays are Enterprise Resource

Planning (ERP) and Customer Relationship Management (CRM) systems. ERP and CRM systems allow access to the data and collect data from diverse operations. The systems can be implemented with various tools, for example, SAP is the most known ERP system, and Microsoft CRM application is the most known CRM system. Other than that, the companies can also implement its own tools for the same purpose.

We looked at the components of Business Intelligence from the perspective of technology, as well as the BI definition of Evelson (2008) and further studies of Martin, Maladhy, Venkatesan (2011), Turban, Sharda, and Delen (2011). Hence, we think that BI includes all the tools / applications and techniques that support decision making, which leads to better understanding of its own business, and therefore achieving its business objectives. We added the most common and recognizable tools and techniques to Figure 3 in order to illustrate these elements in Business Intelligence.

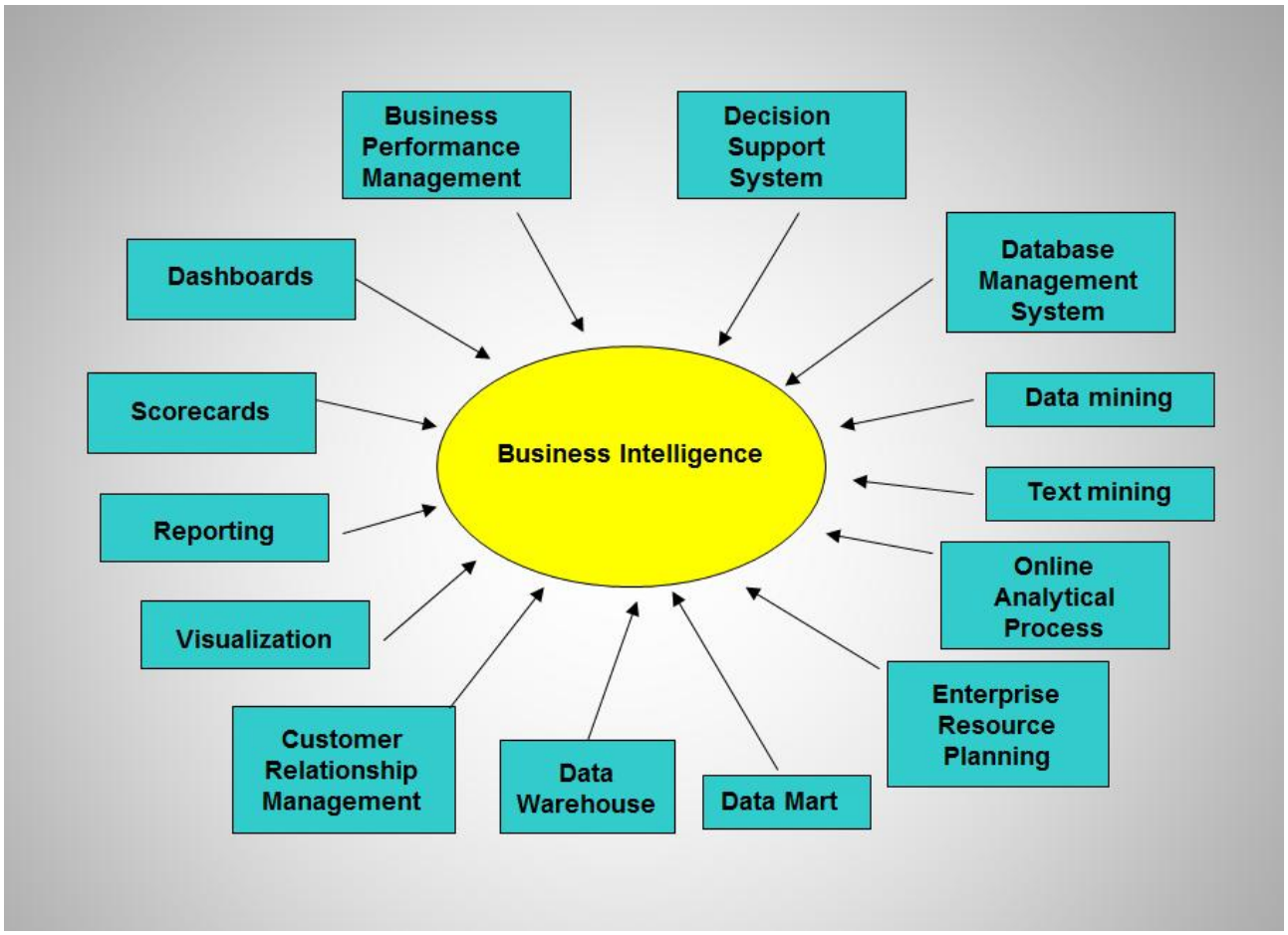

Figure 3: Various tools and techniques for BI

A Business Intelligence System monitors situations and identifies problems and opportunities, using analytic methods. In order to be able to demonstrate the result and findings, reporting plays a major role in Business Intelligence.

On the strategic level of business operation, Business Performance Management and Decision Support System have been well researched before, and several models have been implemented and widely utilized in the industry. Turban, Sharda, and Delen (2011) gave the definition of BPM and DSS, “Business Performance Management (BPM) is an advanced performance measurement and analysis approach that embraces planning and strategy. It refers to the business processes, methodologies, metrics, and technologies used by enterprises to measure, monitor, and manage business performance”. “Decision Support System (DSS) is a conceptual framework for a process of supporting managerial decision making, usually by modelling problems and employing quantitative models for solution analysis. It refers to the information system supporting the company’s decision-making activities due to the rapidly changing business situations.” Models play a major role in DSS and BPM because they are used to describe real decision-making situations and business achievement. These models can be either static or dynamic.

Besides models, some tools have been created to handle and measure BPM as well in the industry, one of the most recognizable tools is the Balanced Scorecard. The Balanced Scorecard (BS) is an innovative and very powerful tool for business performance measurement developed in 1992 by Harvard Business School professor Robert S. Kaplan and management consultant David P. Norton. It can measure whether the corporation has achieved its objectives related to its vision and strategies. The balanced scorecard quantifies the business performance into performance measures and evaluates if the system is balanced between: internal performance and external performance perspectives, short-term objectives and long-term objectives, financial measures and non-financial measures.

2.3 Disadvantages of Business Intelligence System

Business Intelligence System also has disadvantages. Knowing these disadvantages can help companies like Florilla Consulting understand information concerning the BI system, and how to find the most suitable BI system that benefits them. We listed three major disadvantages based on the problems that happened to the most companies based on the studies of Jaglan, Dalal, Dr.S.Srinvivasan (2011) and Turban, Sharda, and Delen (2011).

- Cost and requirement: The cost of establishing business intelligence is always the foremost issue coming into consideration for small and medium size companies due to the high requirements for hardware standard, and the system is expensive for basic business transactions as well, so the cost is often beyond the budget.

- Complexity: It is very complex to setup a suitable BI system to satisfy the company's needs; this involves all the related techniques and processes how to handle the data implementation**;** for example, what data mining technique to use, how to setup an efficient data warehouse.

- Duration of implementation: Many projects failed due to the complexity of the development process, or companies run out of patience to wait for the execution of BI, because the time required for implementing BI is very long, usually 18 months for data warehousing system to completely implement the system.

Even today many companies are not using BI system, not only because they cannot afford it, but also due to its complexity, it requires many professionals to handle the BI system. In addition, many of the companies do not consider BI system to be highly essential.

3. DATA ANALYSIS

3.1 Data Analysis

Data Analysis is the activity of analysing data and finding useful patterns**.** It is the core technology in Business Intelligence, in another word; Business Intelligence System is based on data analysis. Most people may call it Data Mining or Predictive Analytics in both the industry and academia. For example, data analysis is used to detect fraudulent credit card and predict the most likely defaulters in load application process in banking; or to detect sales volumes at each retail locations in order to determine correct inventory levels in logistics; or to demonstrate students how different variables (etc. investment amounts, inventory levels) to influence the final result in a data analysis models at school. In addition, data analysis is particularly useful for CRM customer analytics, banking, health care, retail, and education.

In case of Florilla Consulting, Data Analysis is central in its operation because data analysis, in addition to constituting its core business, also offers the advantages of predicting the result with high accuracy. Likewise, Florilla Consulting can perform a cost-benefit analysis of using data analysis to solve the clients' problems. We will talk more about Cost-Benefit Analysis in the later chapter, but let us look at the key components to handling data analysis first.

3.2 Data Warehouse System

3.2.1 Data Warehouse

Data warehouse (DW) is a repository of current and historical data produced to support decision making. Inmon, W.H. (1992) also stated that Data Warehouse is a subject-oriented, integrated, time-variant, non-volatile collection of data in support of the management's decision-making process.

Han and Kamber (2006) in their book stated that Data Warehouse is defined as a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing, and it is the core of a decision supportive system. Therefore, having proper data warehouse architecture is always a good start of any BI approach towards business.

Data Warehouse system has been studied in the recent decades, and it has become more an important and valuable tool in today's competitive, fast-growing business world. Most data warehouses are built using relational database management systems. Oracle and Microsoft SQL Server are the most commonly used ones, and they all support both client/server and Web-based architectures. One of the important DW factors is the long-term storage of data from multiple sources, but this does not mean that all the data and Extraction Transformation Load (ETL) should be stored and handled in DW repository only. In other words, DW should be a database that is maintained separately from an organization's operational databases.

Meanwhile, in order to minimize redundant usage of the organizational DW system, different users shall be clarified with different roles according to the enterprise's data model. We categorized the things to be noticed based on the overview of BI system Architecture in Figure 2 and the state-of-the-art BI system with four major components which we discussed previously, for example:

- Executives and senior managers shall use standard reports or the data analysts create specialized reports for executives and senior managers
- Data Analysts writing complex SQL queries to load, extract, and compile specific data shall not use existing tools and applications
- Reporting tools and custom applications often need to access the database directly
- Customers shall not interact directly with the relational database, but may receive email, reports or access web pages that extract data from the relational database

3.2.2 Data Warehouse Benefits and Features

Date Warehouse brings many benefits to business, which has been discussed a lot in the industry, so we summarize them as follows based on the studied of Han and Kamber (2006) and Turban, Sharda, and Delen (2011). It:

- Provides a competitive advantage by presenting relevant information to measure performance and make critical adjustments in order to help win other competitors.
- Enhances business productivity by quickly gathering information to describe the organization efficiently and accurately
- Facilitates customer relationship management because it provides a consistent view of customers and items across all lines of business, departments, and markets.
- Brings cost reduction by tracking trends, patterns, and exceptions over long periods in a consistent and reliable manner.

Data Warehouse also has many features and among of them, there are four key features that have been mentioned in the definition of Data Warehouse by Inmon, W.H. (1992). The characteristics of data warehouse have also been described in detail by Han and Kamber (2006) in their book and can be summarized as:

Subject-oriented: Rather than concentrating on the daily operations and business transactions, a data warehouse is used around major subjects, such as customers, suppliers, products, sales, inventories. Data warehouse focuses on modelling and analysis of data for decision makers. Therefore, data warehouse typically provides a simple and concise view around particular subjects by excluding data that are not useful in the decision support process.

Integrated: A data warehouse is usually constructed by integrating multiple sources, such as relational databases, flat files and on-line transaction records. Data integration techniques are applied to ensure the consistency of integration with other data sources.

Time-variant: Data are stored to provide information from the historical perspective. Time is the one important dimension that all data warehouses must support. They detect trends, deviations, and long-term relationships for forecasting and comparisons, leading to decision making.

Non-volatile: A data warehouse is always a physical separate store of data transformed from an application data found in the operational environment, and does not require transaction processing, recovery, and other operations. Hence, a data warehouse requires only two operations in data accessing which are loading the data and accessing the data.

3.3 Data Mart and OLAP

Data Mart is usually smaller and focuses on a particular subject or department; it is a subset of a data warehouse, typically consisting of a single subject area. Based on the description of Data Mart in Turban, Sharda, and Delen (2011)'s book, a data mart can either be dependent or independent.

- A dependent data mart is a subset created directly from the data warehouse; hence it has the advantages of using a consistent data model and provides quality data.
- Independent data mart is a small warehouse designed for a strategic business unit or a department, but its source is not an enterprise data warehouse.

Due to the high cost of data warehouse, many companies use a lower-cost, scaled-down version of a data warehouse, which is referred as independent data mart.

Figure 4 illustrates the relationship between Data Warehouse and Data Mart, where each Data Mart represents a department, and as a whole makes a Data Warehouse.

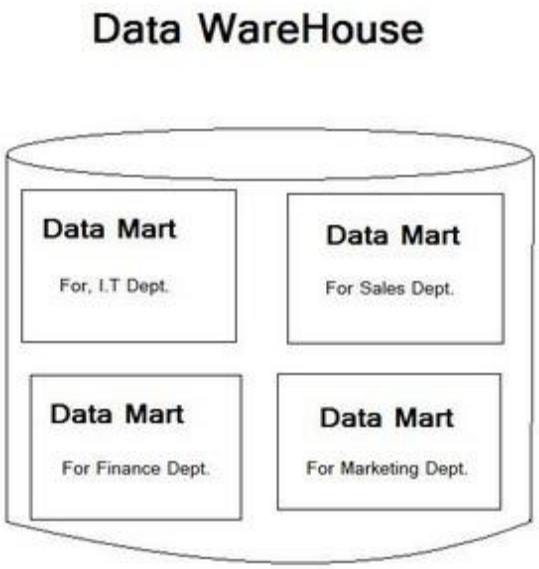

Figure 4: Data Mart and Data Warehouse (Ravi, 2009)

Other than Data Warehouse and Data Mart, another key element of Business Intelligence is Online Analytical Processing. One thing in common about all of them is that they all deal with databases and access data. Prabhu, S (2007) stated in his book that "Online Analytical Processing (OLAP) is a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user."

An OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at different level. OLAP can be considered as a data warehouse tool as well as Data Mining, but the difference is that OLAP supports model-driven analysis and data mining supports data-driven analysis. OLAP database should be different than DW database as well to achieve better performance. In general, OLAP provides a good view of what is happening, but cannot predict what will happen in the future and why it is happening.

Data Warehouse also provides OLAP tools for the interactive analysis of multi-dimensional data, which facilitates effective data generalization and data mining. Many of other DM functions, such as classification, prediction, clustering, and association can be integrated with OLAP operations to enhance the overall mining knowledge.

3.4 BI Architectures with DW, DM, OLAP

Previously, we have talked about the BI architecture and its components in general, now we shall look at how Data Warehouse, Data Mart, and Online Analytical Processing are integrated into one system.

Small and medium-scaled companies are often constrained with the budget on the implementation of an efficient BI architecture, and thus may choose the simplest BI architecture design. Poe et al. (1997) drew the first enterprise data architecture in Figure 5 showing a Data Warehouse feeding Data Marts from his work. The database in Figure 5 could be any type of data source.

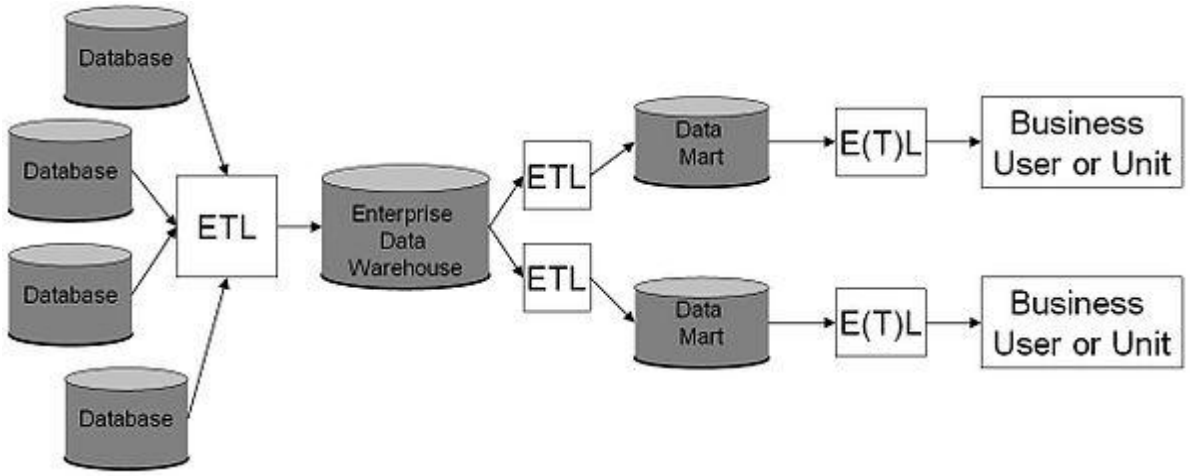

Figure 5: BI Architecture with DW and DM (Wikibooks.org, 2009)

This kind of layout incorporates simple functions, but the performance would not be as good as the more complex ones because all the extraction, transformation and loading of

the data from the database sources are handled in the Enterprise Data Warehouse database, which lowers down the overall performance of the BI system. So OLAP database needs to be separated from the Enterprise DW. It helps to promote the high performance of both systems. An operational database is designed and tuned from known tasks and workloads, such as indexing and using primary keys, searching for particular records, and optimizing queries. While data warehouse queries are often complex and involves the computation of large groups of data at summarized levels, and may require the use of special data organization access and implementation methods based on multi-dimensional views. Processing OLAP queries in operational databases would also degrade the performance of operational tasks. The separation of them is based on the different structures, contents and uses of the data in these two systems, Decision support requires historical data, but operational databases do not typically maintain historical data.

Companies with sufficient budget would prefer a design where an integrated DB feeds a DW, for example in Figure 6 that was proposed by Poe et al. In this case the business users and business unit both extract the data for reporting purposes and also update the data without lowering the overall performance.

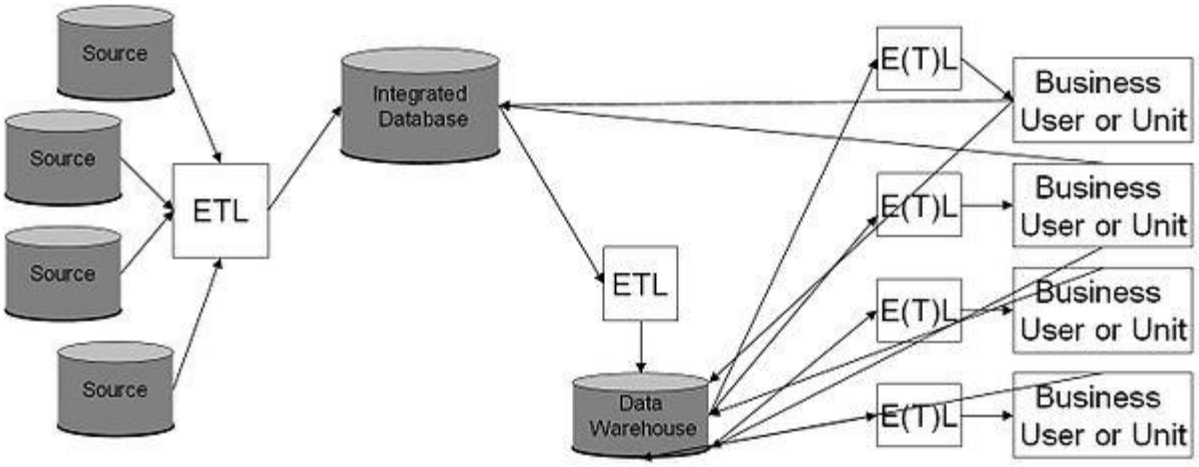

Figure 6: BI Architecture with OLAP and DW (Wikibooks.org, 2009)

Establishing the proper enterprise data warehouse architecture is always a good start for any of the BI approach. This depends on the scale of the company and the financial capability**.** The big players in the market need state-of-the-art BI architecture to sustain its

business requirements. Hence, we think that Figure 7 mentioned in Chaudhuri, Dayal, and Narasayya (2011)'s studies is a good example to represent well-implemented business intelligence architecture. The only difference from the setup in Figure 6 is that the Mid-tier is added, where OLAP server, Data Mining and reporting engines are positioned. This setup would enhance the performance of Data Warehouse to the maximum.

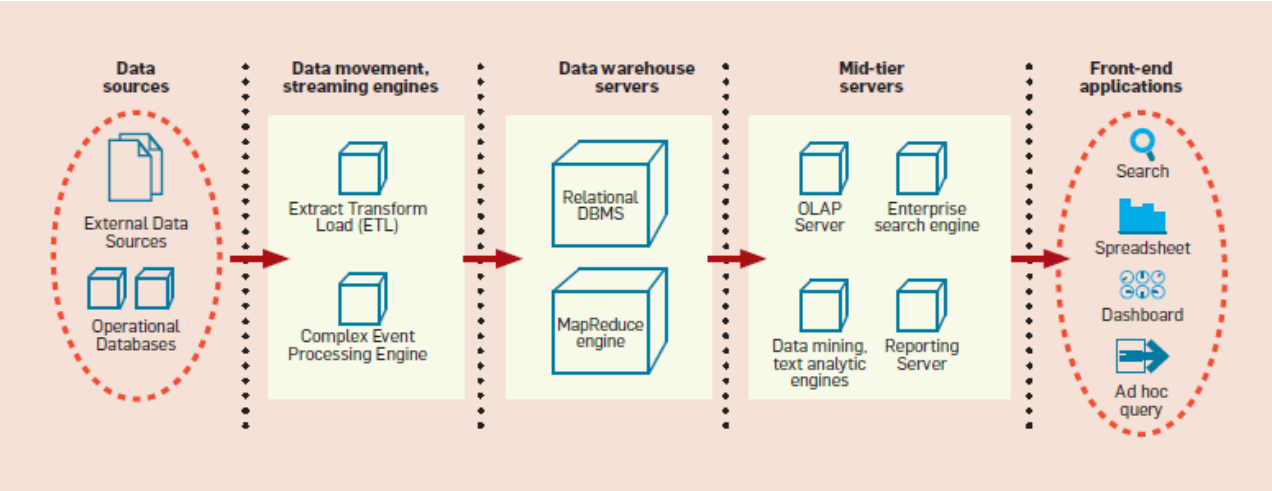

Figure 7: A well-implemented BI Architecture (Chaudhuri, Dayal, and Narasayya, 2011)

3.5 Data Mining

Data Mining is the core technology in Business Intelligence, and it is used to solve business problems, also to gain a better understanding of the customers, own operations, and therefore solve complex organizational problems. Turban, Sharda, and Delen (2011) mentioned the definition of Data Mining, "Data mining is a process that uses statistical, mathematical and artificial intelligence techniques to extract and identify useful information, subsequent knowledge, and patterns from large sets of data."

The data mining environment is usually a Client-Server architecture or Web-based information systems architecture. The miner is often an end user, and the result of DM might be unexpected and requires end users to think creatively throughout the process, and how to interpret the findings. Data mining tools are readily combined with spread

sheets and other software development tools, so that the mined data can be analysed and deployed quickly and easily. Sometimes, parallel processing for data mining is necessary due to large amounts of data and massive search efforts.

A company that effectively leverages data mining tools and technologies may acquire and maintain a strategic competitive advantage. That is, data mining offers companies an indispensable decision-enhancing environment to exploit new opportunities by transforming data into a strategic weapon.

3.5.1 Data Mining Benefits

Data Mining has been widely used to solve business problems, such as customer segmentation and buying patterns, customer relationship management improvement, financial forecasting, credit card fraud detection and credit scoring, inventory level prediction, product recommendation, direct marketing campaign, and thus has been specifically helpful in many areas. Some of them have been proven successfully by the following examples, which were mentioned by Turban, Sharda, and Delen (2011), Han and Kamber (2006) as well as other research work done on the topic of data mining.

Banking:

- Successfully detect fraudulent credit card/visa card and online-banking transactions using data warehouse of the customer transactions.
- Accurately predict the most likely defaulters in load application process.
- Maximize banking customer value by selling them products and services that they are most likely to buy.
- Accurately forecast the cash flow in each branch and ATMs.

Retailing, manufacturing and logistics:

- Accurately detect sales volumes at each retail locations in order to determine correct inventory levels.

- Identify anomalies in production system, and predict the most likely to-be-happened bottleneck of the manufacturing process to optimize the manufacturing capacity.
- Accurately forecast the consumption levels of each product type based on seasons to optimize logistics and maximize sales.

Stock trading:

- Predict when and how much certain fund and bond prices will change.
- Forecast the range and direction of stock fluctuations
- Evaluate the effect of particular issues and events on overall market movements

Health care:

- Forecast the level and the time of demand at different service locations to optimally allocate resources
- Explore new cost-benefit relationships between different treatments to develop more effective strategies
- Identify the people who have no health insurance and the reasons of this phenomenon

KD nuggets (http://www.kdnuggets.com) is one of the data mining community's top resource websites. In 2011, KD nuggets conducted an online pool to collect data on the usage of data mining in industries / fields and compared with data they collected from 2010, Figure 8 shows the results of the comparison in each field. CRM consumer analytics, Banking, and Health care / HR are still the top four fields that applied data mining, whereas the usage of data mining in Education has increased largely.

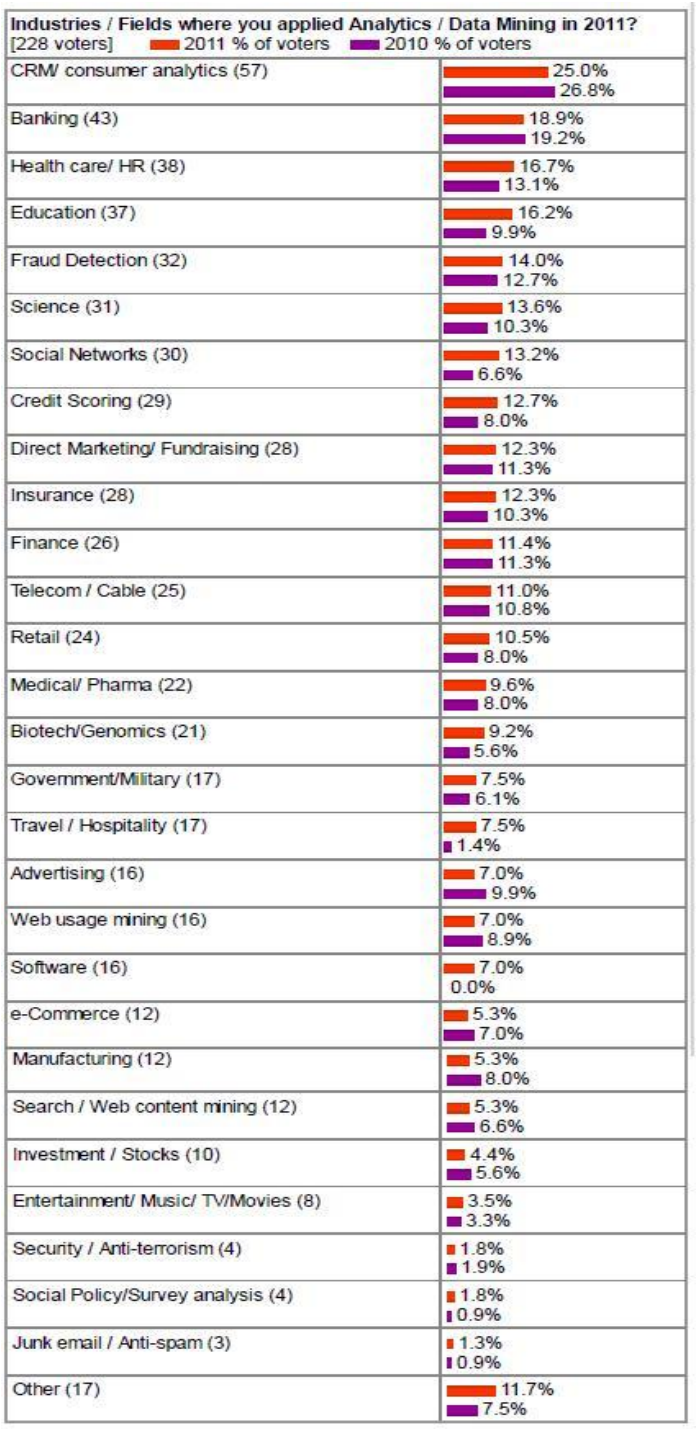

Figure 8: Industries / Fields where Data Mining is applied in 2011 (KD nuggets, 2011)

In most cases, data mining refers to tabular data mining, which means that the data has been processed to have tabular data format as an input before going into data module. There are also non-tabular data mining, such as text mining, multimedia mining, web mining, etc. This is because data for data mining comes in many different forms, for

example, some standard database format or business information in SQL; some type of computer file (.txt, .xls) typed by human; information recorded automatically by tools such as error log file; or business information from internet websites.

Other than the most often used tabular data mining, text mining, and web mining are becoming more popular. We give definitions of such as bellows based on Turban, Sharda, and Delen (2011)'s book:

Text mining is the semi-automated process of extracting patterns from large amounts of unstructured data sources. Text mining is the same as data mining in that it has the same purpose and uses the same processes, but with text mining the input to the process is a collection of unstructured data files such as Word documents, PDF files, and XML files. Text mining is one of the fastest growing branches of the business intelligence field.

Web mining is the process of discovering intrinsic relationships, for example interesting and useful information from web data, which are expressed in the form of textual, linkage, or usage information, Web mining consists of Web content mining, Web structure mining, and Web usage mining.

Text and Web mining are emerging as critical components of the next generation of business intelligence tools enabling companies to compete successfully. Hence, Florilla Consulting could use text mining and web mining to better understand their customers by analysing their feedbacks on web forms, blogs, and wikis.

KD nuggets conducted an online pool to collect what data types have been analysed / mined in the past 12 months until June, 2011. Figure 9 gives a good overall of the data mining trend, which can help Florilla Consulting know their future focus.

Data types analyzed/mined in the past 12 months [206 voters total]	
table data (fixed n. columns) (143)	69.4%
time series (86)	41.7%
itemsets / transactions (67)	32.5%
text (free-form) (53)	25.7%
anonymized data (45)	21.8%
location/geo/mobile data (40)	19.4%
other (29)	14.1%
social network data (26)	12.6%
email (22)	10.7%
web content (21)	10.2%
web clickstream (18)	8.7%
images / video (14)	6.8%
XML data (10)	4.9%
music / audio (7)	3.4%

Figure 9: Data types analysed in the past 12 months (KD nuggets, 2011)

After all, Data Mining is a complex process**,** and in the industry Data Mining tasks heavily depends on data mining professionals to provide solutions. In addition, data mining neither is a “black box” process in which the data miner simply builds a data mining model and watches as meaningful information appears**;** nor does it guarantee the behaviour of future data through the analysis of historical data. Instead, data mining is a guidance tool, used to provide insight into the trends inherent in historical information.

3.5.2 Data Mining Processes

Data mining process is defined with stages from selection, pre-processing, to transformation, mining, and at last result in interpretation or evaluation. There are many variations of the data mining process, for example, Cross-Industry Standard Process for Data mining and SEMMA. Companies use different processes and models for the different DM tasks.

A data mining study must be viewed as a process that follows a standardized methodology rather than as a set of automated software tools and techniques.

3.5.2.1 CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) is the most popular standardized process, which was first established in the middle of 1990s by European Consortium of Companies, which included Integral Solutions Ltd, NCR, DaimlerChrysler, and OHRA. Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler) (2000) published the first version of CRISP-DM that gave step-by-step data mining guide.

Figure 3 illustrates the process of CRISP-DM. In the diagram, each step does not necessary lead to the next step but rather it is an iterative process as specified.

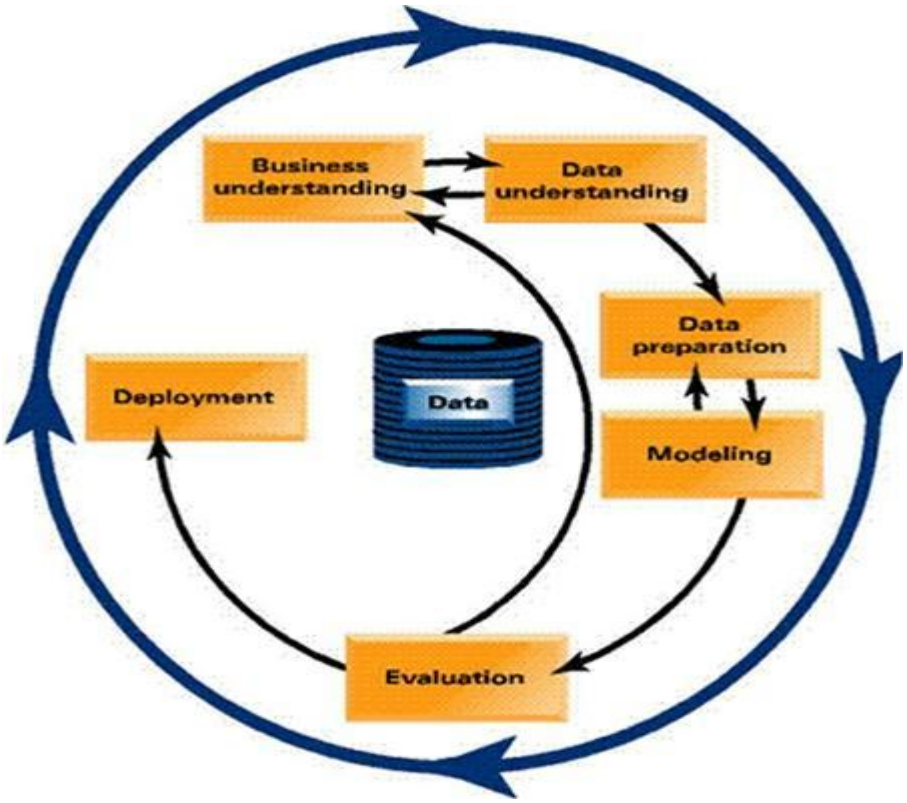

Figure 10: CRISP – DM diagram (Tom Khabaza, 2010)

Hence, we summarize CRISP-DM steps as below based on the handbook of CRISP-DM 1.0 and the studies of Turban, Sharda, and Delen (2011)

Step 1: Business Understanding

The first step is to determine the goal of this data mining study. A specification of the business objectives including the success criteria should be defined first, such as “Why is the inventory level always unexpected at certain site?” or “What kind of customers should the company sell certain product to?” That is, a project plan to find such knowledge should be developed first and the data analysts, who collect data, analyse data, report the findings should be assigned too.

Meanwhile, a budget to support such a study should be estimated as well by for example the project manager.

Step 2: Data Understanding

The second step is to familiarize with the initial data collection, explore the data and identify the data quality. After knowing the business objectives, the main activity of the DM process is to identify the relevant data from large data sources. Solving the DM questions can be addressed using querying, reporting, and visualization. It is crucial to understand the data mining task clearly and concisely so that the most relevant data can be identified. A careful identification and selection of data sources and the most relevant variables can make it easier for DM algorithms to quickly discover the useful knowledge patterns. The last thing to check before the next step is to examine the quality of the data whether the data is completed or not.

Step 3: Data Preparation

Data preparation is the most crucial one and it consumes the most time in a DM project due to incomplete and inconsistent raw data from the data source. Every activity after the initial raw data and before the final data construction is handled here.

Data preparation is essential to any successful data mining study. This is because good data leads to good information, and good information leads to good decisions. Unfortunately, in real life the data that act as the data source is very often not in a ready format as an input due to cases such as**:** information that cannot be obtained; a malfunction of the equipment or tool that used to record the data; carelessness of the people who handle the data source; or a data collection form with additional fields that were added after the previous data had been collected. So, knowing these facts can prevent from enormous work of data preparation and therefore reduce the error of input data.

Four main steps are used in the process of data preparation:

- Data Consolidation**:** collect and filter the data, then integrate and unify the data by SQL queries or Microsoft Excel.
- Data Cleaning**:** handle the missing values in the data, identify the outliers in the data and eliminate erroneous data.
- Data Transformation**:** derive more information variables from the existing ones using mathematical functions, or using data normalization, discretisation, binarisation to reduce the range of the values or convert the numeric variables into discrete representations.
- Data Reduction is to reduce the number of attributes and records using for example principle component analysis, decision tree induction, and random sampling.

Step4: Model Building

The model-building step includes the assessment and comparative analysis of the various models based on the data mining objective. Likewise**,** parameters are calibrated to the optimal values. There is no a universally known as the best algorithm or methods for a data mining task, so data mining tasks can use a variety of data mining algorithm and methods.

Modelling can be associated with the decision trees, neural networks, association analysis, regression, clustering, and time sequence analysis, and in model deployment phase, different techniques may be applied in each step. The accuracy of an algorithm depends on the nature of the data. For example, a decision tree algorithm is usually a very good choice for any classifications. However, if the relationships among attributes are complicated, a neural network may perform better. A good approach is to build multiple models using different algorithms, and then compare the accuracy of these models. Even with a single algorithm, you can tune the parameter settings to optimize the model accuracy. Last but not the least, data mining is an exploratory process, and it often takes a few iterations before finding the right model.

Identification of a model's variables is critical as well as the relationships among the variables. The components of a model for decision support usually are result variables, decision variables, uncontrollable variables and intermediate result variables. The results of decisions are determined based on the decision made (the values of the decision variables), the factors that cannot be controlled by the decision maker, and the relationships among the variables. The modelling process involves identifying the variables and relationships among them. Solving a model determines the values of these and the result variables. Result variables reflect the level of effectiveness of a system, which indicate how well the system performs in achieving its goals. Decision variables are the alternatives of the action, for example how much money to invest or people to allocate on one project. Uncontrollable variables are the factors that affect the result variables but cannot be controlled by the decision makers. Intermediate result variables reflect intermediate outcomes in a mathematic model, for example, employee salaries can affect

employee satisfaction, and hence can determine the productivity level, which influence the final result.

Step5: Testing and Evaluation

This step is to assess and evaluate the developed models for the accuracy and generality, and see if the selected model meets the business objectives. A key objective is to access if there is any important business issue that has not been considered sufficiently.

Meanwhile, the model should be tested in a real application or a real case to see if the time and budget constraints permit.

The testing and evaluation is a critical and challenging task. Determining the business value from discovered knowledge patterns requires the effort from data analysts, business analysts, and decision makers in order to proper interpret knowledge patterns. Tables and visualization would be a good means to demonstrate the findings to customers.

Step 6: Deployment

The last step is to organize and present the knowledge gained from the data mining task in a way that the end users understand and benefit from it. It is actually the customer, not the data analyst, who carries out the deployment steps. It is also important to maintain activities for the deployed models so that the data mining results become part of the day-to-day business and its environment due to the constantly changing business pace.

3.5.2.2 SEMMA

“Sample, explore, modify, model and assess (SEMMA)” is another well-known methodology developed by the SAS Institute. The SAS Institute is an organization, which is one of the largest privately-held corporations in the software business. SAS website also

explained SEMMA process and its related product SAS Enterprise Miner. Figure 11 demonstrates the SEMMA process:

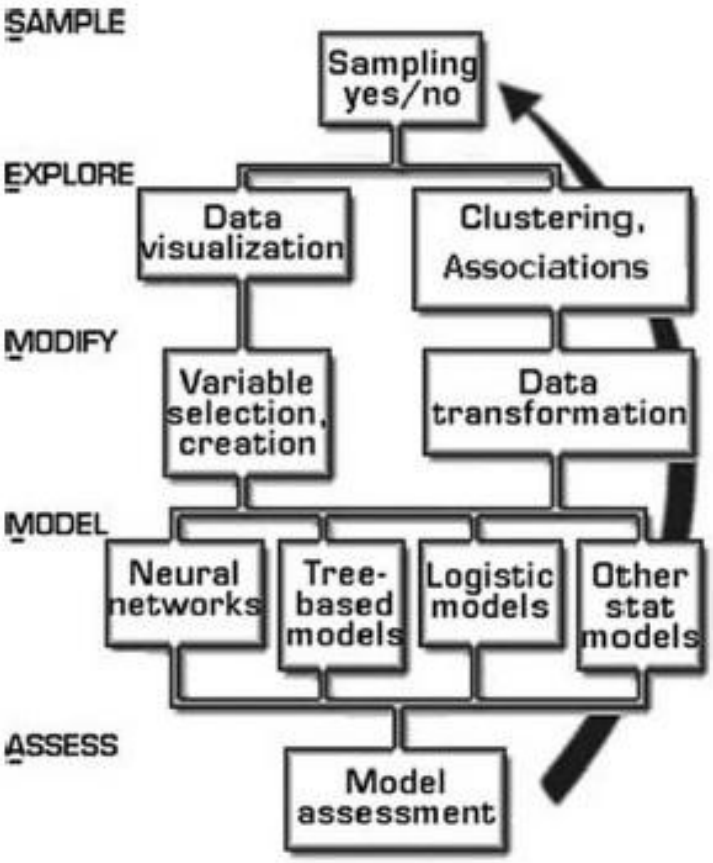

Figure 11: SEMMA diagram (Data Prix, 2010)

Beginning with a statistically representative sample of the data, SEMMA makes it easy to apply exploratory statistical and visualization techniques that select and transform the most significant predictive variables, model the variables to predict outcomes, and confirm a model's accuracy. We describe the process in detail as below:

- Sample: Generate a representative sample of the data
- Explore: Visualization and basic description of the data
- Modify: Select variables, transform variable representations
- Model: Use variety of statistical and machine learning models
- Assess: Evaluate the accuracy and usefulness of the models

By assessing the outcome of each stage in the SEMMA process, the model developer can determine how to model new questions raised by the previous results, and thus proceed back to the exploration phase for additional refinement of the data**.** We see that SEMMA is driven by a highly iterative experimentation cycle.

The main difference between CRISP-DM and SEMMA is that CRISP-DM takes a more comprehensive approach including understanding of the business and relevant data pertaining to the data mining projects, whereas SEMMA implicitly assumes that the data mining project's goals and objectives alone with the appropriate data sources have already been identified and understood.

3.5.3 Data Mining Methods

Data mining studies are central to Florilla Consulting. Currently, there are several methods of performing data mining studies, including prediction, classification, regression, clustering, and association. Most data mining software tools use more than just one algorithm or technique for each of these methods.

To be able to conduct data mining successfully, the data analysts in Florilla Consulting need to know what those algorithms can do and how to make use of them sequentially, and how to set the parameters as well.

In general, when we start the data mining process, we look at a dataset which is called instances, and each of the dataset comprises of a number of variables with values, which is called attributes in data mining. There are two types of data, which are treated in different ways.

One type is a specially designated attribute and the aim is to use the given data to predict the value of the attribute that has not been seen yet. This kind of data is called labelled data, and data mining using labelled data is known as supervised learning. Another type is the data that does not have any special attribute is called unlabelled, and data mining of

unlabelled data is known as unsupervised learning. Supervised learning includes classification and numerical prediction, while unsupervised learning includes association rules and clustering. The aim is simply to extract the most information we could from the data available.

To give an overall of which data mining algorithms have been used in the industry of 2011, we use Figure 12 to demonstrate the result, which was collected from an online pool on KD nuggets website.

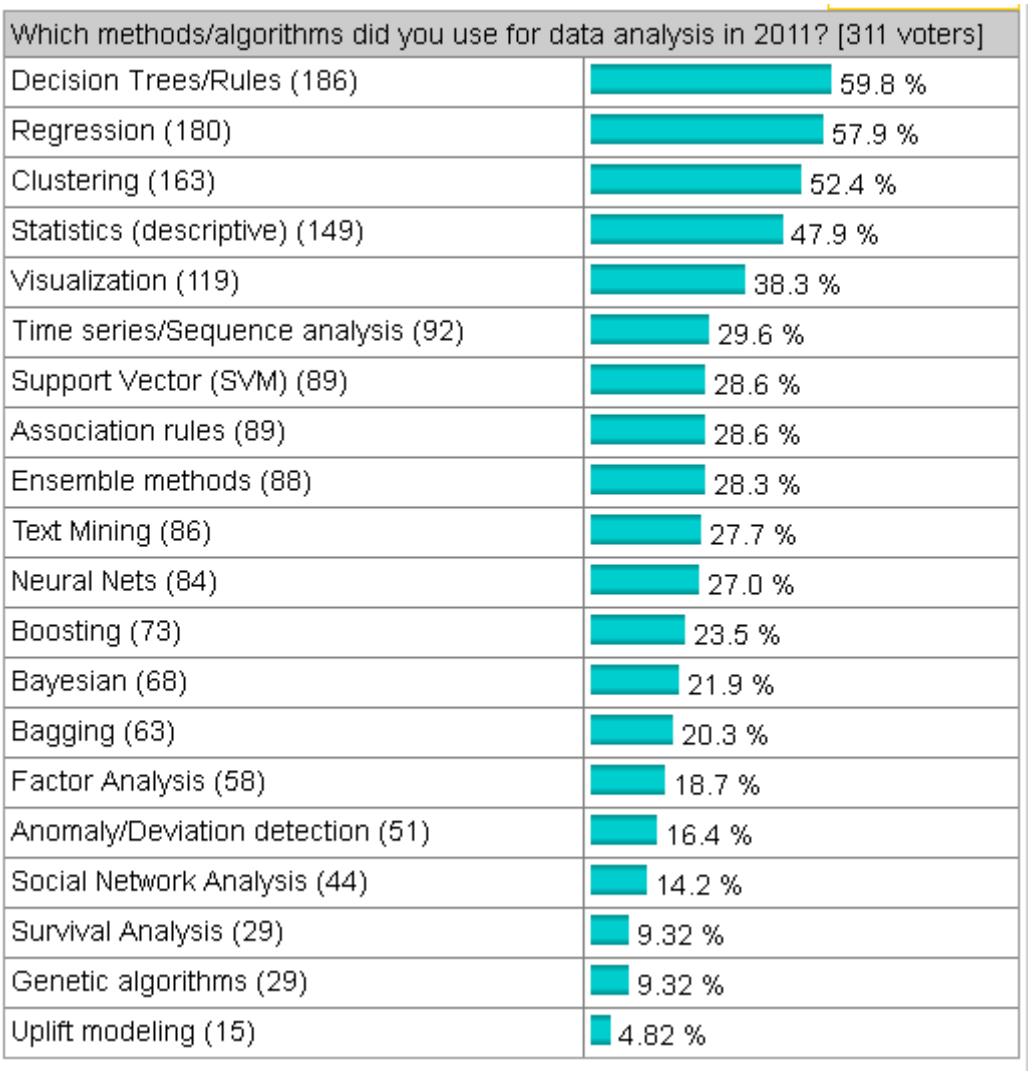

Figure 12: Data analysis algorithms used in 2011 (KD nuggets, 2011)

Many of the same data mining algorithms can be used in different data mining methods, so we categorize data mining methods into five main categories based on studies of Han and Kamber (2006), Turban, Sharda, and Delen (2011), Bramer (2007), and Wu, Kumar, et al.(2007) among the others.

3.5.3.1 Prediction

Prediction is commonly referred as telling the future, and is associated with the term forecasting. It should be noted that prediction is largely experience based, while forecasting is data and model based. Prediction encompasses the identification of distribution trends based on the available data.

Forecasting takes sequences of numbers indicating a series of values through time as input then imputes the future values of those series using a variety of machine-learning and statistical techniques that deal with seasonality, trending, and noisiness of data.

Prediction and classification are often discussed to have similar meanings, but are also discussed as two forms of data analysis used to extract models describing important data classes or to predict future data trends. In order to make the structure easier to understand, classification and numerical prediction are often put under the umbrella term of prediction. While classification predicts categorical labels / classes, numerical prediction models continuous valued functions.

Predictive accuracy, computational speed, robustness, scalability, and interpretability are the five criteria for the evaluation of classification and prediction methods. In particular, robustness is defined as the model's ability to make reasonably accurate predictions, given noisy data or data with missing and erroneous values.

3.4.3.2 Classification

Classification is the most frequently used data mining method**,** and it occurs frequently in everyday life. It predicts categorical class labels and constructs a model to describe a set

of predetermined classes by the class label attribute based on the historical data stored in a database.

Many decision-making tasks can be formulated as classification problems, for example

- Customers who are likely to buy or not buy a particular product in a certain market
- Clients who are eligible to get loan from the bank
- Students who are likely to obtain or not obtain a scholarship from a university

Classification Techniques

There are a number of techniques that are used for classification modelling, for example, Decision Tree Analysis, Statistical Analysis, Neural Networks, Bayesian Classifiers, Case-based Reasoning, Genetic Algorithms, but decision tree analysis is the most commonly used technique among of others, which can also been seen in Figure 12.

Decision Tree Analysis

A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions.

There are many good facts about decision tree analysis in Data Mining, such as:

- Faster learning speed relatively compared to other classification methods
- Convertible to simple and easy to understand classification rules
- Can use SQL queries for accessing databases
- Comparable classification accuracy with other methods

Figure 13 is a good example to demonstrate a data mining model using decision tree analysis in financial risk analysis; the financial staff can follow the tree branches based on the conditions to decide whether it is too risky to give loan to the customer or not based on the customer's income and credit history.

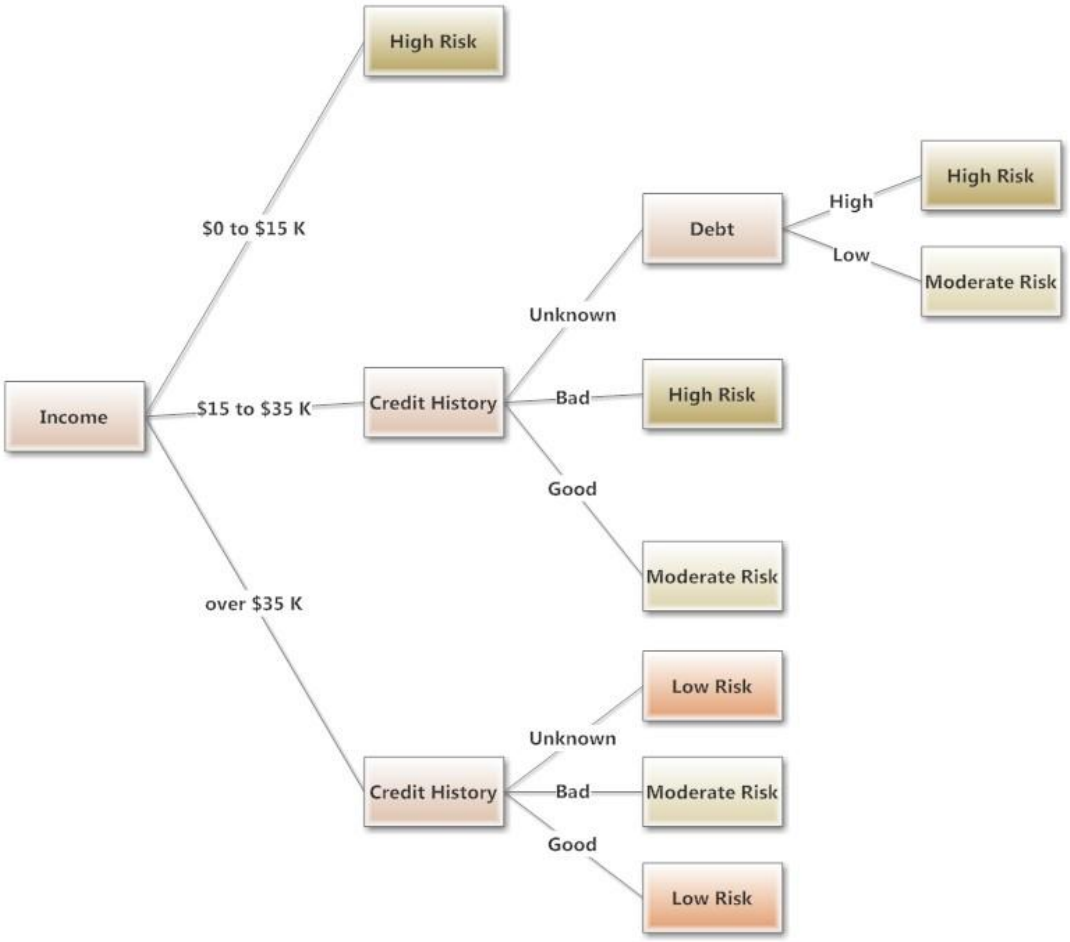

Figure 13: Decision Tree Method used in financial risk analysis (Smart Draw, 2012)

In addition to the other classification techniques, there is a quite particular technique called Case-based reasoning, which we think it is beneficial to use for Florilla Consulting.

Case-based Reasoning

Case-based Reasoning (CBR) uses a database of problem solutions to solve new problems. It is based on the premise that new problems are often similar to previously encountered problems, and the successful solutions in the past may be of use in solving the problem that arises at the current situation.

When creating a new case, CBR checks if any previous identical case exists. If it does, then the solution to the case is returned. If no identical case is found, then it searches for

cases having similar components to the new case. In fact, the CBR uses background knowledge and problem-solving strategies in order to propose a feasible combined solution, and the benefit of accuracy and efficiency evolves as the number of stored cases becomes large. When the number increases, the DBR becomes more intelligent.

For the companies that handle large amount of data mining tasks, it is very efficient and convenient to have a case-based DM platform or repository. The platform should consist of at least three parts:

- a data storage repository that stores all the pass data
- a DM methods (algorithms) storage repository that stores the different algorithms of data mining, for example decision tree analysis, Bayesian model
- a DM model storage repository that stores the different models which have been used in the past DM tasks

3.4.3.3 Numerical Prediction

Numerical Prediction is sometimes called regression analysis; it is a statistical methodology that is most often used for numeric prediction. It is generally considered as “statistics” rather than “data mining”. For many data mining tools that do not include a regression module is that it is hard to market regression as a leading edge technology.

Numerical prediction has techniques like Linear Regression, Nonlinear Regression, Neural Network, and K-mean. As a matter of fact, K-mean is the most common method used for numerical prediction, which is also used commonly in cluster analysis.

3.4.3.4 Cluster Analysis

Cluster analysis is an essential data mining method for classifying items, events, or concepts into common groups called clusters. It is concerned with grouping together objects that are similar to each other and dissimilar to the objects that belong to the other clusters. Custer analysis is an exploratory data analysis tool for solving classification problems.

Cluster analysis has been used extensively for fraud detection (both credit card and e-commerce fraud) and market segmentation of customers in CRM systems.

Grouping similar objects together has some major benefits in many fields, for example:

- In business strategy to find the countries whose economies and cultures are similar
- In finance to find customers of companies that have similar financial performance
- In marketing to find the clusters of customers that have similar buying behaviours

Clustering can be used to determine groups as well, but there is a significant difference between clustering and classification. Classification learns the function between the characteristics of variables, like independent variables and output variables**,** through a supervised leaning process where both types of variables are presented to the algorithm. In clustering, on the other hand, the output is learned through an unsupervised learning process where only the input variables are presented to the algorithm. Clustering algorithms, then, use one or more heuristics to discover natural groupings of objects instead of enforcing the learning process based on supervising mechanism.

Cluster Analysis Techniques

K-means

The k-means algorithm is the most referenced clustering algorithms, k stands for the predetermined number of clusters. It has its roots in traditional statistical analysis. The algorithm assigns each data point such as customer, event, object, to the clusters whose centre is the nearest. The centre is calculated as the average of all points in the clusters. Figure 14 demonstrates a simply K-means method to find the centre of income and age.

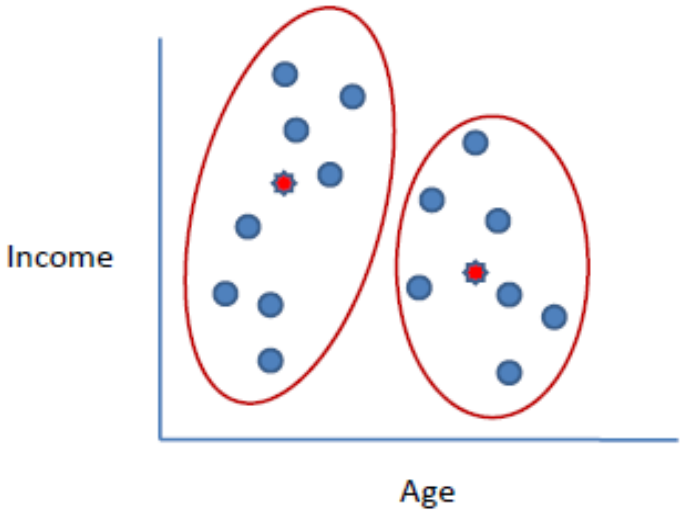

Figure 14: K-means (Sayad, 2010)

The procedure would be to choose the number of clusters first, i.e. randomly generate k random points as initial cluster centres, then assign each point to the nearest centre, and recomputed the new cluster centres. At last, repeat the last two steps until come convergence criterion is met, which means the assignment of points to clusters becomes stable.

Other than K-means, there are also other similar techniques, such as Neural networks, Funny Logic, Genetic Algorithms.

3.4.3.5 Association Rule Mining

Association is to find the commonly co-occurring groupings of data in general. Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. It is widely used for market basket or transaction data analysis. The primary aim of association rule mining is to examine the contents of the database and find rules in the data.

Two common used derivatives of the association rule mining are link analysis and sequence mining. With link analysis, the linkage among many objects of interest is

discovered automatically, such as the link between milk and bread. Usually customers buy milk also buy bread. With sequence mining, relationships are examined in terms of their order of occurrence to identify associations over time. In the context of the retail industry, association rule mining is often called the Market-Basket Analysis.

Apriori Algorithm is the most common used association rule mining, others are known as FP-Growth, OneR, ZeroR, Eclat.

3.5.4 Data Mining Tools

Many software vendors provide powerful data mining tools for the customers, for example, Microsoft SQL Server DM, ORACLE DM, SAS/SAS Enterprise Miner, RapidMiner Enterprise Edition, SPSS PASW Modeler (formerly Clementine), Microsoft Excel, MATLAB.

In addition to these commercial tools, several free data mining software tools are available as well, for example, RapidMiner Community Edition, and Weka.

The main difference between commercial tools and free tools is the efficiency of computation. The same data mining task involving a large dataset may take longer time to complete with free software, and it may not even be feasible in some case, for example, crashing due to running out of computer memory.

Figure 15 demonstrates most of the popular data mining tools that have been used for a real project in 2011. The questionnaire was conducted among Enterprise and Academy with a response rate of 1103 participants via an online pool on KD nuggets website. 43% of them used only commercial software, 32% used only free software, and 25% both. There has been a slightly decrease with RapidMiner, but it is still the most popular data mining software tool. Nevertheless, RapidMiner, R, and Excel are still on the top three, with SAS remaining the top commercial tool.

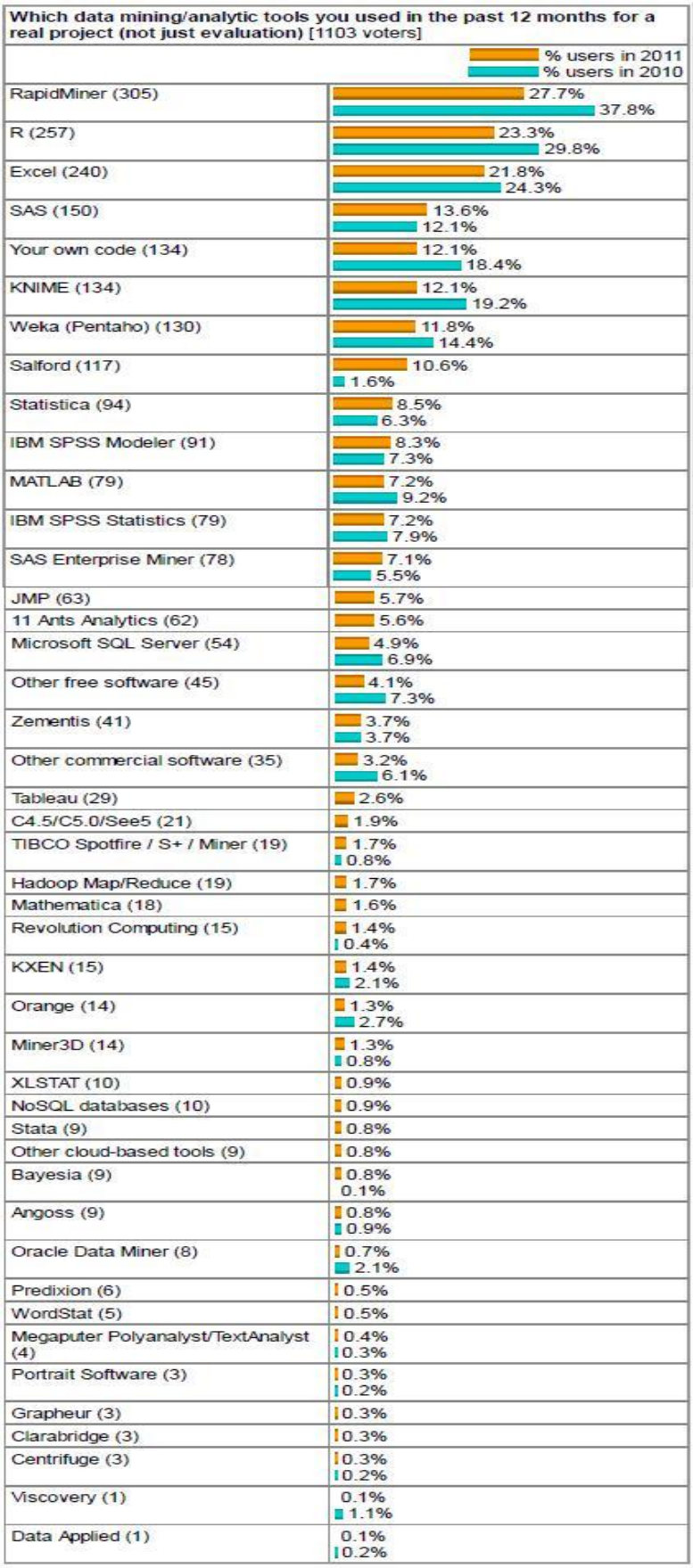

Figure 15: Popular data mining tools used for a real project in 2011 (KD nuggets, 2011)

3.5.4.1 Commercial Tools

3.5.4.1.1 Microsoft

Microsoft's SQL Server 2008 is one of the most popular business intelligence capable suites for data mining studies. The data and the models are stored in the same relational database environment so that the performance is faster and the model management is considerably easier.

Microsoft Business Intelligence Solution is a full suite of application servers, clients and developers, fully integrated with the Microsoft Office System, which delivers desktop business information into a central and integrated menu. Microsoft Business Intelligence solution includes back-end servers such as the SQL Server, and front-end applications like Excel, Word, and Visio. It is designed for interoperability with the data that exists in almost any database source of the company, like Oracle, IBM DB2.

Microsoft's SQL Server 2008 provides an integrated environment for creating and working with the data mining models, called Business Intelligence Development Studio. The studio includes data mining algorithms and tools that make it easy to build a comprehensive solution for different projects. Figure 16 demonstrates its own data mining process, which is similar to CRISP-DM.

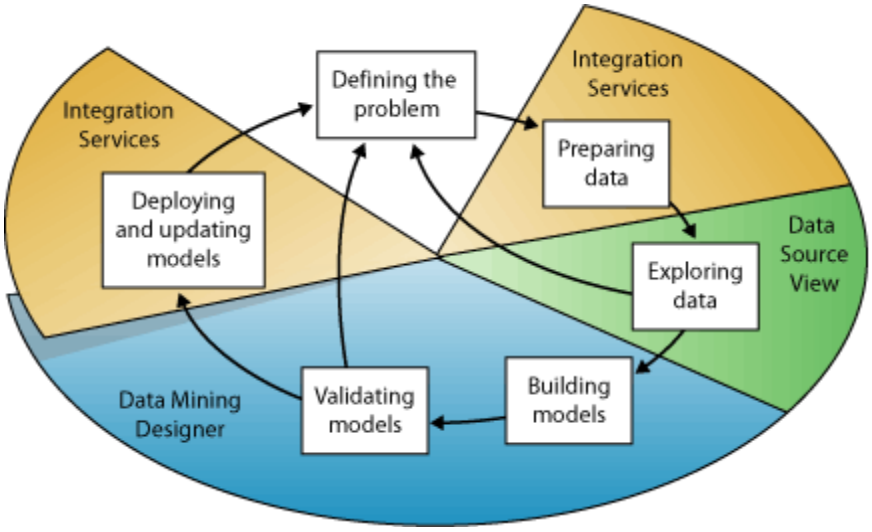

Figure 16: Microsoft Data Mining Process (Microsoft.com , 2005)

3.5.4.1.2 Oracle Enterprise Miner software

Oracle Data Mining is part of the Oracle Relational Database Management System which can be enabled to handle data mining with Oracle databases. It contains several data mining algorithms for prediction, classification, and association among others and helps analytics to create, manage and operate deployment of data mining models inside the database environment.

3.5.4.1.3 RapidMiner Enterprise Edition

RapidMiner is an open-source system for data mining. It is available as a stand-alone application for data analysis and as a data mining engine for the integration into own products. It combines data integration, analytical ETL, data analysis, and reporting in one single suite**.** At the same time, it provides powerful graphical user interface for design of analysis processes and supports hundreds of data loading, data transformation, data modelling, and data visualization methods.

RapidMiner provides solutions that can be applied on various fields and even combines the tasks of structured and unstructured analysis that is necessary for sentiment analysis. Sentiment analysis is the automatic identification of positive and negative emotions, opinions, and evaluations from any text.

Besides RapidMiner, another product called RapidAnalytics can act as a team server for RapidMiner. With this setup, the analysts can collaborate very efficiently and can even share computing resources of a large central server. Apart from that, RapidAnalytics can also be used to present interactive reports to the customers when it becomes necessary.

3.5.4.2 Free Tools

RapidMiner Community Edition is the same product as RapidMiner Enterprise Edition but with much less features. That is, only for the purpose of testing with a small scale of data. This tool also lists the available data mining models within, but is not stable, which increase the risk of data mining failure, and turn out to be more expensive. Other than that, it cannot provide any type of support or guarantee for any implementation without an Enterprise Subscription contract.

3.5.5 Data Mining Integration with Databases or Data Warehouse System

Previously we have studied that good system architecture will facilitate the data mining system to make the best use of the software environment, accomplish data mining tasks in an efficient and timely manner, interoperate and exchange information with other information systems, and easily adapt to users' diverse requirements. But there is a critical question, "how much integration should data mining be with the database system or data warehouse system?"

Turban, Sharda, and Delen (2011) mentioned several methods of integrating database and data warehouse, which we think it is good to discuss here in order to help Florilla

Consulting understand the setup of data mining with data warehouse system. If a DM system works as a stand-alone system or is embedded in a software application, which does not communicate with any DB or DW systems, this is called no-coupling scheme. It mainly focuses on developing effective and efficient algorithms for mining the available data sets. However, when a DM system works in an environment that requires it to communicate with the other information system components in DM and DW systems, these schemes are known as loose coupling, semi-tight coupling, and tight coupling.

No coupling implies that DM does not utilize any function of DB or DW. In other words, the DM tool usually reads data from a particular source such as a file system and processes data using some DM algorithms, then stores the result in another file. It is the most economical way in terms of investment, but it has several drawbacks as well.

First of all, data in DB and DW systems tend to be well organized, indexed, cleaned, integrated, so that finding the task-relevant, high-quality data becomes an easy task. DB and DW systems also provide great flexibility and efficiency at storing, organizing, accessing, and processing data plus scalable algorithms and implemented data structures. Without using the DB/DW system, DM may spend a large amount of time on finding, collecting, cleaning, and transforming data. Another drawback is that without using the DB/DW system, DM has to use other tools to extract data, which makes it difficult to integrate such a system into an information processing environment.

Loose coupling means that a DM system uses some facilities of a DB/DW system to fetch data from a data repository and perform data mining, then store the results either in a file or in DB/DW system.

Loose coupling is better than no coupling because it fetches any portion of the data stored in the DB/DW system by using query processing and indexing, but it does not explore data structures and query optimization methods provided by the DB/DW system. So in this case it is difficult to achieve high scalability and good performance with large datasets.

Semi-tight coupling means efficient implementations of some essential DM features such as sorting, indexing, aggregation, histogram analysis, multi-way join, and some statistical

measures (sum, count, max min…). Moreover, some frequently used intermediate mining results can be pre-computed efficiently and stored in the DB/DW system and these very same mining results enhance the performance of a DM system.

Tight coupling means a DM is smoothly integrated into a DB/DW system. Data mining queries and functions are optimized and all of them are integrated together as one information system with multiple functionalities. This approach is highly desirable because it facilitates efficient implementation of the data mining function, high system performance, and integrated information processing environment.

From a cost-benefit point of view, no coupling has the minimum investment, but has quite many drawbacks and may not benefit business performance so much; Loose coupling is better than no coupling, but not efficient; tight coupling is highly desirable, but its investment is the highest, its implementation is nontrivial and more research is needed; While on the contrary, semi-tight is a compromise between loose and tight coupling, its investment is often affordable usually, and the benefit to business performance is evident**.** Hence, we think that Semi-tight integration is a very good option for Florilla Consulting.

3.5.6 Data Mining Mistakes

We collected some of the common problems and mistakes that analytics should always avoid when performing data mining based on the studies of Turban, Sharda, and Delen (2011), and John Elder (2005), which we hope Florilla Consulting would pay attention to:

- Asking the wrong questions, it is important to have the right project goal
- Selecting the wrong problem for data mining
- Not paying attention to the sponsor opinions about data mining and only listen to the data
- Lack of proper data
- Not having sufficient time for data preparation

- Looking only at aggregated results and not at individual records; some of the interesting records should be highlighted
- Not keeping full track of the data mining procedures and results
- Ignoring suspicious findings and quickly moving on
- Repeating the same data mining algorithms without a clear thought
- Believing the data too much and thinking that everything has been told
- Believing the best model
- Measuring results differently from the way sponsor measures them

3.6 Data Analysis Synthesis

So far we have introduced the core technology Data Analysis in Business Intelligence, and also have discussed the major components in Business Intelligence to handling data analysis. Data Analysis is considered as Data Mining sometimes, it is based on the technology with Data Warehouse, OLAP, and Data Mart as a whole to achieve remarkable business performance, based on the fundamental data infrastructure, especially when Business Performance Management and Decision Support System are used in daily operation. Meanwhile, Data Mining can also be used separately without applying Data Warehouse in data mining project. This would not require high standard of Business Intelligence IT infrastructure, and it is mainly used for customer projects.

We discussed not only the benefits of using Data Warehouse, differences between Data Warehouse and Data Mart, but also different methods of deploying Business Intelligence Architectures with Data Warehouse, Data Mart, and OLAP, as well as the benefits and flaws of having certain architecture, and then we recommended a well-implemented Business Intelligence Architecture to Florilla Consulting.

Then we emphasized on the topic Data Mining, its benefits, how it shall be integrated with Databases and Data Warehouse, and the mistakes that shall be concerned. We especially emphasized on Data Mining processes and most popular commercial and free tools that have been used widely in Data Mining industry. In addition, we also categorized the

current Data Mining methods into five different categories and discussed about them and also related techniques, mainly for Florilla Consulting data analysts to handle data mining projects.

4. METHODOLOGY

There are many methodologies to analyzing a technology. Preferred by Florilla Consulting, as also confirmed by my instructor who is working in Florilla Consulting, we decided to apply Cost-Benefit Analysis as methodology to analyze the data analysis framework especially the Data Mining tools that we propose to Florilla Consulting. Therefore, we introduce the concept of Cost-Benefit Analysis as below.

Cost-Benefit Analysis

Benefit-Cost Analysis Centre at the University of Washington's Daniel J. Evans School of Public Affairs gave the definition of Cost-Benefit Analysis from an economic perspective as: "Cost-benefit Analysis (CBA), also known as benefit-cost analysis (BCA), aims to inform the decision-making process with specific types of information, namely measures in monetary terms of willingness to pay for a change by those who will benefit from it, and the willingness to accept the change by those who will lose from it"

Fuguitt Diana, Wilcox Shanton J (1999) also gave the definition of Cost-Benefit Analysis from a societal perspective as:
"Cost-Benefit Analysis is a useful approach to assess whether decisions or choices that affect the use of scarce resources promote efficiency. Considering a specific policy and relevant alternatives, the analysis involves systematic identification of policy consequences, followed by valuation of social benefits and costs and then application of the appropriate decision criterion."

There are many definitions of Cost-benefit Analysis from different perspectives, thus we think that CBA can be described as an economic decision-making approach to compare the benefits and costs of a decision. It collects all the positive and negative factors and quantifies them to find out whether it is beneficial to make the decision.

Cost-Benefit Analysis is one of the most widely used methods for deciding whether a project investment is a good use of a project resource. Hence, Cost-Benefit Analysis

technique is being applied to this research. For example, in the case of having Business Intelligence for Florilla Consulting, we will analyse its costs and benefits, especially for data mining tools, which will be summarized in the next chapter.

5. DATA ANALYSIS FRAMEWORK FOR FLORILLA CONSULTING

Based on the literature review we have studied so far, we propose a data analysis framework for Florilla Consulting to conduct their daily business. This framework consists of Business Intelligence architecture, data mining process, and data mining tools, which we think the most suitable ones for Florilla Consulting, but it also depends on the financial situation of Florilla Consulting and its customer requirement and environment.

In the following sub-chapters we will examine some of the Business Intelligence key components and core technology to demonstrate the benefits of having them for Florilla Consulting in this research.

5.1 BI Architecture

The Business Intelligence Architecture of any company highly depends on the size and the business goals of the company. Based on the literature review previously, we think that Business Performance Management and Decision Support System both are good options to choose in order to measure, monitor, and manage business performance, as well as help managers to make decisions quickly. However this would require the setup of Data Warehouse, and we think that the Business Intelligence Architecture in Figure 6 with integrated OLAP is strongly recommendable if Florilla Consulting aims to monitor its business performance and wants to quickly act to the changes in the competitive market.
On top of that, the Balanced Scorecard is a very handle and recommendable tool to measure Florilla Consulting business performance as well.

However if Florilla Consulting is currently at the beginning phase of initiating its business and constrained by the budget, then data warehouse may be quite expensive to build and maintain in the beginning.

5.2 Data Mining Process

Irrespective of whether data warehouse is chosen or not, the data mining process must be executed in Florilla Consulting. That is, the data mining process is something that cannot be avoided in achieving trustful results, as the data mining process must be reliable and repeatable by employees with little data mining background.

CRISP-DM is definitely a mature and widely-used data mining process and can be used as a standard data mining process for Florilla Consulting. This is not only based on the literature we have done on Data Mining Process leading to the conclusion of using CRISP-DM, but also because CRISP-DM provides a uniform framework for guidelines and experience documentation and is flexible to different data and business problems as well.

We developed a model for Florilla Consulting data mining process based on the original CRISP-DM module with description for each step.

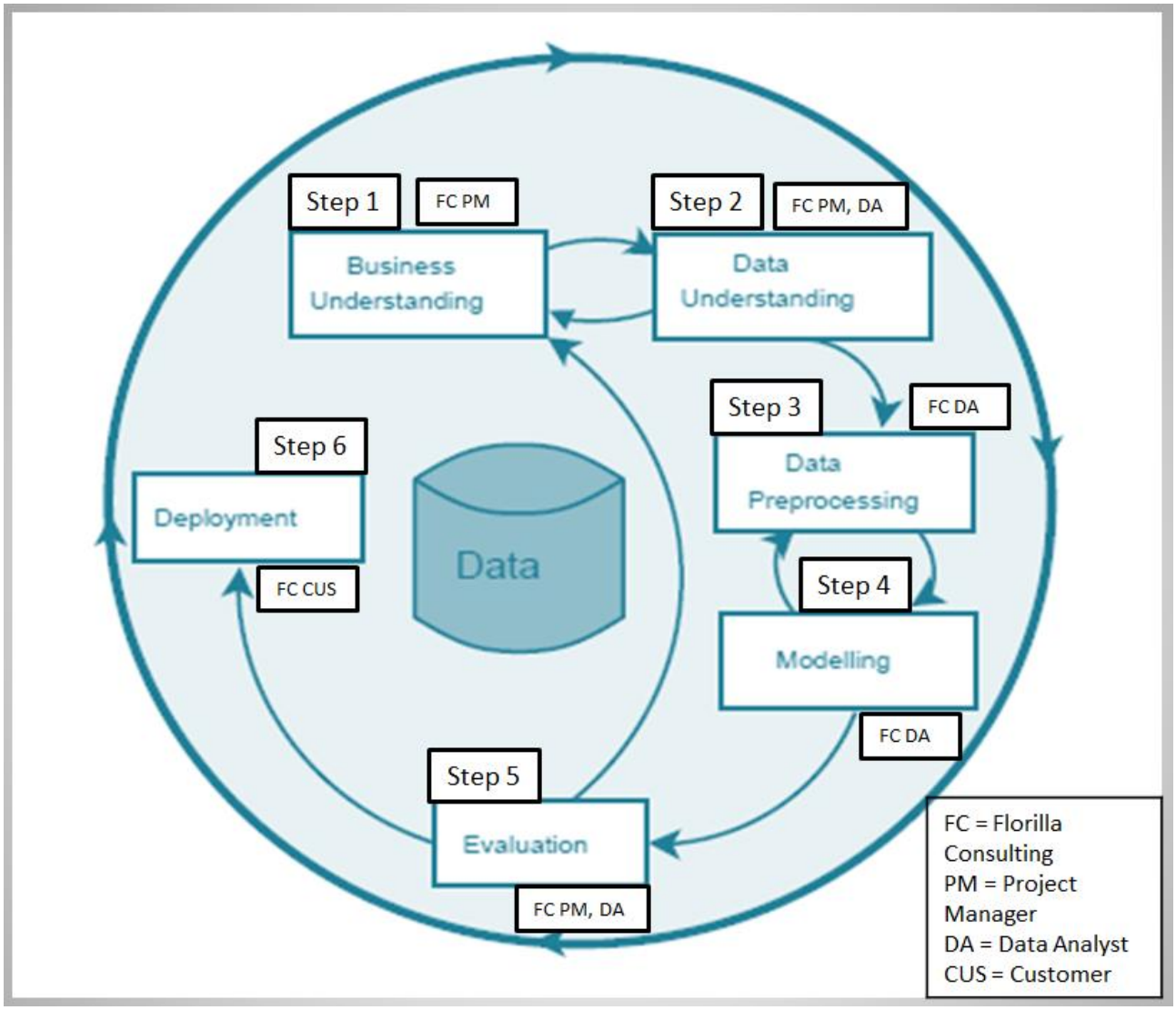

Figure 17: Model for Florilla Consulting Data Mining Process

Step One:

The manager of the data mining project in Florilla Consulting should be able to understand the objectives clearly and involved with the data analysts to explore the feasibility of the project.

Step Two:

The senior data analyst takes a big role in this step in order to understand the metadata of the data source and the target database. The senior data analyst needs to work closely with the project manager in case that the understood data may not be subject to business

needs. If the senior data analyst notices that the collected data may not be able to reach the business objectives, either the raw data or the objectives in step one shall be adjusted.

Step Three:

Once the goal and the data are clearly understood, the data pre-processing can start. There are several ways to handle this.

Microsoft Excel is a very commonly-used and handy tool for this task, some missing or invalid data can be spotted by the data analysts manually via a pre-validation tool with validation rules established to find all the invalid data.

Microsoft's SQL Server 2008 also provides an integration service to handle the pre-processing of data. This is a more advance tool that provides more features, but requires the data analysts at Florilla Consulting to create the design of the process to filter out the invalid data. It is a good option if Microsoft's SQL Server 2008 is chosen as the option**.**

Step Four:

Modelling is actually the process that controls what comes in and what comes out. The senior data analysts need to define the variables and relationships beforehand, for example, the decision variables, result variables, and uncontrollable variables. Doing so can help the data analysts to find the best model that tunes the most correct result. The data analysis can also decide what data mining algorithm to use in the model, but remember "One size does not fit all" so that no single method has been found to be superior over all others for all data sets.

Due to the wide data mining filed of Florilla Consulting, we developed table 1 based on the literature review on CRISP-DM to demonstrate some examples of how to find different variables for modelling in general.

Area	Decision Variables	Result Variables	Uncontrollable Variables and Parameters
Financial investment	Investment amounts and other alternatives	Operating profit, Earnings per share, Liquidity level	Inflation rate, Financial crisis, Prime rate, Competition
Marketing	Advertising budge, marketing location	Market share, Customer satisfaction	Customer's income, Competitor's actions
Manufacturing	Productivity, Inventory levels, Compensation programs	Total cost, Quality level, Employee satisfaction	Machine capacity, Technology, Raw materials prices
Accounting	Use of computers Audit schedule	Data processing cost, Error rate	Computer technology, Tax rates, Legal requirements
Services	Budget, Staffing levels	Customer satisfaction, Brand image	Demand for services

Table 1: DM model variables in different field

Step three and four are close each other, in case the model cannot be found accurately, which means that there might be some flaws in the data pre-processing step. So the data analysis may have to go back to previous step and investigate the issue to be able to find the accurate model for data mining.

Step Five:

Once the model is ready, the most important factor to take into account for Florilla Consulting is to check if the selected model meets the business objectives. If the model has been tested with good results, this can be used as a model for future in the projects, the only details that need to be changed are the variables. Otherwise, the data analyst

may have to go back to step one, and see if the business objectives have been understood clearly.

Step Six:
The last step is the time to deploy the data mining process to the customer sites and see if the customer accepts the result.

As far as for what Florilla Consulting is concerned, the deployed models become the company's assets and the data mining results become part of the day-to-day business due to the constantly changing of business. Hence, on top of these steps, it would be very beneficial for Florilla Consulting to use Case-based Reasoning that we have studied in the literature review, to develop a knowledge pool for future development.

5.3 Data Mining Tools

The data mining tools that are appropriate for Florilla Consulting should be chosen based on the Company's needs and financial status. The most expensive one may not be the best option, and vice versa. Florilla Consulting should choose the most suitable one for its business operation, which not only depends on the budget, but also depends on the data mining field and the customer's requirement.

If Florilla Consulting wants to have its own data mining infrastructure, this would require not only the database itself, but also the data mining tool; however, if the deployment is at customer site, then the data mining tool would depend on the setup of the customer's database infrastructure, for example some companies only use Oracle database or Microsoft database. No matter what tool Florilla Consulting decides to use, the data mining deployment should always follow the data mining process strictly.

Let us take a look at the options for data mining tasks:

Microsoft SQL Server Enterprise would cost approximately 20,000 -22,000 € for the server and its license**.** It includes features like Data Warehousing, Advanced Data Mining and Data Integration, Enterprise Data Management, etc. SQL server also provides a light version for the single user license usage which does not require any SQL server setup and implementation, and the cost is approximately 230€ per usage, but this may not provide enough features for the daily operation of Florilla Consulting unless the data analyst of the Company knows exactly what to get from it.

Oracle provides the enterprise version, which has high costs. As a matter of fact, the minimum costs for the implementation would total up to 28,000€ - 32,000€. Oracle has been considered as one of the most expensive database providers. There is also a case, when the customer has Oracle database setup, and the data mining would work better with the Oracle data mining features. So this highly depends on the customer's database infrastructure. In case of Florilla Consulting, if the company currently possesses business analysts who are familiar with the Oracle technology, then the cost problem is partially reduced by offering consulting services to the customer's onsite.

Microsoft Excel is probably the most common Microsoft Office tool that all companies are using for daily business. The basic features of Excel always have data sorting and data filtering, which are extremely helpful for processing data. It also has an additional feature for data mining, but it requires Microsoft SQL Server Analysis Service in addition.

RapidMiner Standard Enterprise Edition offers full support for technical as well as analytical problems for up to five users, so setting it up will not take as much effort as SQL Server Data Mining. Besides that**,** RapidMiner is also open to Florilla Consulting's own codes, which allows the combination of these codes to the power of a flexible framework. This is a good feature of the software that makes it much more flexible in terms of the data sources than the case with SQL Server DM, which highly necessitates the SQL server. For example, Florilla Consulting may save a lot of costs by using MySQL instead of the SQL Server.

After the marketing research we have done and tacit knowledge for different Database and Data Mining tools licenses and specifications, we made Table 2 to give an overall of the mentioned data mining tools for Florilla Consulting with a cost-benefit analysis. Due to the limitation of the final performance, we cannot quantify the benefit in numbers. Therefore we use categories (LOW, MEDI, HIGH, and VERY HIGH) to represent the performance at a similar level. The scale of this cost-benefit analysis is 5 persons minimum due to the practical requirement and software license for using the software.

	Cost	Benefit	Additional Concerns
Microsoft SQL Server	20,000 - 22,000 €	HIGH	• Beneficial with MS SQL database • Flexible with Microsoft Excel, RapidMiner • High costs for database maintenance and update
Microsoft SQL singer user license usage	1150€ (230€ x 5ppl)	LOW	• Beneficial with MS SQL database • Insufficient for data mining
Oracle Server	28,000€ - 32,000€	HIGH	• Beneficial with Oracle database • High costs for database maintenance and update
Microsoft Excel	1000€ (200€ x5ppl)	VERYHIGH	• Extremely useful • Commonly used in ICT companies • Flexible with Microsoft Excel, RapidMiner, Oracle, and company's own code
RapidMiner Standard Enterprise Edition	7100-7300 €	VERYHIGH	• Less effort to set up • Open source • works with MySQL • Low costs for maintenance
Florilla Consulting own code	17500€ (3500€ x 5m)	MEDIUM	• Hard to predict how much time to develop a sufficient tool for data mining models • Require senior data analyst for implementation

Table 2: An overall of data mining tools for Florilla Consulting with a cost-benefit analysis

After a Cost-benefit Analysis view, it is clear and easy to see that RapidMiner Standard Enterprise Edition is the most recommendable tool for data mining due to all concerns from different aspects. RapidMiner Standard Enterprise Edition not only works with any of the other databases and tools, but is also open source for Florilla Consulting with a very reasonable annual software license cost. On top of that, it also works with the world most popular open source and free database – MySQL, which lowers costs on data mining projects significantly.

We also think that Microsoft Excel is the most beneficial tool for Florilla Consulting daily business due to the lowest cost and the functionality it provides, even though it does not perform the full function of data mining. Nevertheless, it still is the most recommendable tool as well.

Microsoft SQL Server and Oracle Server are on the same level in terms of costs and performance because they both are commercial tools. Hence, the choice highly depends on the customer data mining requirement and the data analyst skill of Florilla Consulting, but the benefit is high for Florilla Consulting to handle complicated data mining projects. There is a particular case that if Florilla Consulting prefers using RapidMiner but also requires Microsoft database, then Microsoft SQL singer user license usage becomes a good option, but in general, choosing Microsoft SQL singer user license usage is not recommendable for handling data mining projects.

The only uncertain option for data mining task is to use company's own code. This highly depends on the data analysts' skill, and therefore makes the project manager difficult to plan the time schedule to develop a sufficient tool for data mining project. Hence, Florilla Consulting shall carefully choose for this option.

6. CONCLUSION AND DISCUSSION

So far, we have employed a wide range of academic research as a foundation for this study where the most well-known literature is explored particularly in the fields of Business Intelligence and Data Analysis. We first introduced Business Intelligence and its components, how companies can benefit from it, and also the disadvantages of using Business Intelligence System.

Then, we introduced the core technology Data Analysis in Business Intelligence, and discussed the major components like Data Warehouse, Data Mart, and OLAP in Business Intelligence to handling data analysis. We also analyzed these components and how they should be integrated as a whole in Business Intelligence System to achieve remarkable business performance. In addition, we recommended a well-implemented Business Intelligence Architecture to Florilla Consulting.

Later, we emphasized on the topic Data Mining, which is central in the operation of Florilla Consulting. We discussed its benefits and integration with Databases and Data Warehouse in different scenarios. We especially focused on Data Mining processes and most popular tools that have been used widely in Data Mining industry. In addition, we also discussed Data Mining methods and its techniques and then categorized them mainly for Florilla Consulting data analysts to handle data mining projects.

Additionally, we also looked at the various evaluation methodologies and discussed with the representative of Florilla Consulting and decided to use Cost-Benefit Analysis, in order to assess the appropriateness of the case in question.

On top of the theories provided from the academic point of view and in line with the industry trend, we have, then, tailored a data analysis framework including a specific data mining process plus the related tools for Florilla Consulting based on our literature. In addition, both this process and the tools are a result of discussion and confirmation with the Company representatives.

This research is studied in the mainstream of data analysis in the industry from the general level of Business Intelligence System and Architecture, as well as its components, benefits and disadvantages, to the details of Data Warehouse system, Data Mining processes, tools and integrations that have been use mostly in the industry.

However the new appearing Web-based Business Intelligence approaches have not been discussed due to the immaturity of the technology. Nowadays, some small companies also require cheap, lightweight architectures and tools like hardware and software to provide online data analysis, and hence Web Warehousing is a recent approach that merges data warehousing and BI systems with web technologies. A future research would be beneficial to explore the usage of Web Warehousing from a Cost-Benefit Analysis perspective for small and middle-scaled companies.

Due to the scale of the data mining project and limitation of the try-out, we cannot test our results with the customer data using different data mining tools. The software license for database setup and data mining tools limited the further study of this research. In addition, implementing a full process of data mining task takes a huge amount of time and requires expertise.

Finally, the data analysis framework that we have built as a result of this research is beneficial to other similar consulting companies in data analysis field as well. A future research would be interesting to deploy the same data mining project with different data mining tools and see which one performs the best according to speed.

REFERENCE

1. A.Martin, D.Maladhy, Dr.V.Prasanna Venkatesan. (2011) “A Framework for Business Intelligence Application Using Ontological Classification”, *International Journal of Engineering Science and Technology, Vol*. 3 No.2 2 Feb

2. Ales Popovic, Tomaz Turk, Jurij Jaklic. (2010) Conceptual Model of Business Value of Business Intelligence Systems. Available at: http://hrcak.srce.hr/file/81743
   Access date: 18.05.2012

3. Andrew Pavlo, Erik Paulson, Alexander Rasin, Daniel J. Abadi, David J. DeWitt, Samuel Madden, Machael Stonebraker. (2009) A Comparison of Approaches to Large-Scale Data Analysis. Available at:
   http://database.cs.brown.edu/sigmod09/benchmarks-sigmod09.pdf
   Access date: 12.05.2012

4. Benefit-Cost Analysis Centre, Daniel J. Evans School of Public Affairs, University of Washington’s, Available at:
   http://evans.washington.edu/research/centers/benefit-cost-analysis/about
   Access date: 20.05.2012

5. Boris Evelson, (2008) “Topioc Overview: Business Intelligence”, Report for business process professionals, 21,November

6. Christian Schieder, Peter Gluchowski. (2011) Toward A Consolidated Research Model for Understanding Business Intelligence Success. *ECIS 2011 Proceedings.* Paper 205. Available at: http://aisel.aisnet.org/ecis2011/205
   Access date: 10.05.2012

7. Data Prix. (2010) Figure 11, Enterprise Miner's SEMMA process, Available at: http://www.dataprix.net/en/blogs/respinosamilla/theory-data-mining
   Access date: 22.05.2012

8. Dr. Varun Kumar, Anupama Chadha. (2011) “An Empirical Study of the Applications of Data Mining Techniques in Higher Education”, *International Journal of Advanced Computer Science and Applications, Vol. 2, No. 3, March 2011*

9. Efraim Turban, Ramesh Sharda, Dursun Delen. (2011) Decision Support and Business Intelligence Systems, 9th edition

10. Frans Coenen. (2004) Data Mining: Past, Present, Future, *The knowledge Engineering Review, Vol.*00:0, 1-24

11. Fuguitt Diana , Wilcox Shanton J. (1999) Cost-Benefit Analysis for Public Sector Decision Makers

12. Gerrit Lahrmann, Frederik Marx, Robert Winter, Felix Wortmann. (2010) Business Intelligence Maturity Models: An Overview, Available at: http://hrcak.srce.hr/file/81745
Access date: 25.04.2012

13. Gregory S. Nelson. (2010) Business Intelligence 2.0: Are we there yet? Available at : http://support.sas.com/resources/papers/proceedings10/040-2010.pdf
Access date: 29.04.2012

14. Inmon W.H., (1996) “Building the Data Warehouse”, Second Edition, J.Wiley and Sons, New York

15. Jiawei Han, Micheline Kamber. (2006*) Data Mining: Concepts and Techniques*, 2nd edition

16. John Elder (2005), Top 10 Data Mining Mistakes – and how to avoid them, Elder Research, Inc, Available at: http://www.salford-systems.com/doc/elder.pdf
Access date: 10.05.2012

17. Kaplan, Robert S, David P. Norton. (1992) “The Balanced Scorecard: Measures that drive performance” , *Harvard Business Review:* 71-79

18. KD nuggets. (Dec 2011) Figure 8, Industries / Fields where Data Mining is applied in 2011, Available at: http://www.kdnuggets.com/polls/2011/industries-applied-anaytics-data-mining.html Access date: 22.05.2012

19. KD nuggets. (Jun 2011) Figure 9, Data types analyzed in the past 12 months, Available at: http://www.kdnuggets.com/polls/2011/data-types-analyzed-mined.html Access date: 22.05.2012

20. KD nuggets. (Nov 2011) Figure12, Data analysis algorithms used in 2011, Available at: http://www.kdnuggets.com/polls/2011/algorithms-analytics-data-mining.html Access date: 22.05.2012

21. KD nuggets. (Dec 2011) Figure15, Popular data mining tools used for a real project in 2011, Available at: http://www.kdnuggets.com/2011/05/tools-used-analytics-data-mining.html Access date: 22.05.2012

22. Lida Xu, Li Zend, Zongzhi Shi, Qing He, Maoguang Wang. (2007) “Research on business intelligence in enterprise computing enrivonment”, *Systems, Man and Cybernetics, 2007, ISIC. IEEE International Conference*, 3270-3275.

23. Max Bramer. (2007) “Principles of Data Mining”, 1st edition

24. Melanie Hilario, Nada Lavrac, Joost N. Kok. (2010) “Third-Generation Data Mining: Towards Service-Oriented Knowledge Discovery”, SoKD’10, Sep 24, Spain, Available at : http://cui.unige.ch/~hilario/sokd10/sokd10-proceedings.pdf Access date: 04.05.2012

25. Melissa Hardy, Alan Bryman. (2004) *Handbook of data analysis*, 1st edition

26. Microsoft's SQL Server 2008, Available at: http://www.microsoft.com/sqlserver
Access date: 12.05.2012

27. Microsoft.com , (2005) Figure16, Microsoft Data Mining Process, Available at:
http://msdn.microsoft.com/en-us/library/ms174949%28v=sql.90%29.aspx
Access date: 22.05.2012

28. Muntean, Mihaela, Tarnaveanu, Diana, Paul, Anca. (October 23, 2010) "BI Approach for Business Performance". *5th WSEAS Conference on Economy and Management Transformation, 2010*. Available at SSRN:
http://ssrn.com/abstract=1732190
Access date: 22.04.2012

29. Oksana Grabova, Jerome Darmont, Jean-Hugues Chauchat, Iryna Zolotaryova. (2010) "Business Intelligence for Small and Middle-Sized Entreprises". *Journal of SIGMOD Record 39, 2 (2010)* 39-50

30. Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rudiger Wirth, (2000) "CRISP-DM 1.0 Step-by-step data mining guide". Available at
ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf
Access date: 20.04.2012

31. Prabhu, S. (2007) Data Mining and Warehousing

32. Prof.Mihaela I. Muntean, Liviu Gabriel Cabau. (2011) Business intelligence approach in a Business performance context. *Paper presented at MPRA Paper No.* 29914, posted 28, March

33. Rapid Miner, Available at: http://rapid-i.com/
Access date: 22.03.2012

34. Ravi. (2009) Figure 4, Data Mart and Data Warehouse, Available at: http://allaboutdatawarehouse
Access date: 22.05.2012

35. Sayad. (2010) Figure14, K-means, Available at:
http://chem-eng.utoronto.ca/~datamining/dmc/clustering_kmeans.htm
Access date: 22.05.2012

36. Scholz, Patrick, Schieder, Christian, Kurze, Christian, Gluchowski, Peter, Boehringer, Martin. (2010) "Benefits and challenges of business intelligence adoption in small and medium-sized enterprises", *Journal of 18th European Conference on Information Systems,* ECIS2010-0252.R1

37. SEMMA, SAS Enterprise Miner, Available at:
http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html
Access date: 10.05.2012

38. Smart Draw. (2012) Figure 13, Decision Tree Method used in Banking Credit System, Available at:
http://www.smartdraw.com/examples/view/financial+risk+analysis+decision+tree/
Access date: 22.05.2012

39. Surajit Chaudhuri, Umeshwar Dayal, Vivek Narasayya. (2011), Figure 7, A well-implemented BI Architecture , "An overview of Business Intelligence Technology", *Magazine of Communications of the ACM*, volume 54 issue 8, August

40. Tom Khabaza. (2010) Figure 10, CRISP – DM diagram, Available at: http://khabaza.codimension.net/index_files/9laws.htm
Access date: 22.05.2012

41. Umesh Kumar Pandey, S. Pal. (2011) “Data Mining: A prediction of performer or underperformer using classification”*, International Journal of Computer Science and Information Technologies, Vol. 2 (2*), 686-690

42. Veronica Stefan, Mircea Duica, Marius Coman, Velentin Radu. (2010) Enterprise Performance Management with Business Intelligence Solution. Available at: http://www.wseas.us/e-library/conferences/2010/Cambridge/ICBA/ICBA-32.pdf
Access date: 22.04.2012

43. Vivek Jaglan, Surjeet Dalal, Dr.S.Srinvivasan. (2011) Improving performance of business intelligence through case based reasoning, *international Journal of Engineering Science and Technology*, Vol.3 No.4 April 2011.

44. Wikibooks.org. (2009) Figure 5, BI Architecture with DW and DM. Available at : http://en.wikibooks.org/wiki/File:Data_Warehouse_Feeding_Data_Mart.jpg
Access date: 22.05.2012

45. Wikibooks.org. (2009) Figure 6. BI Architecture with OLAP and DW, Available at: http://en.wikibooks.org/wiki/File:Integrated_Database_Feeding_a_Data_Warehouse.jpg
Access date: 22.05.2012

46. Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, Dan Steinberg. (2007) Top 10 algorithms in data mining. Available at:
http://www.cs.uvm.edu/~icdm/algorithms/10Algorithms-08.pdf
Access date: 04.05.2012