Huyen Phan

AN EXPLORATION OF BIG DATA AND ANALYTICS SOFTWARE

Thesis

CENTRIA UNIVERSITY OF APPLIED SCIENCES

Industrial Management

May 2020

Centria
UNIVERSITY OF APPLIED SCIENCES

ABSTRACT

Centria University of Applied Sciences	**Date** May 2020	**Author/s** Huyen Phan
Degree programme Industrial Management		
Name of thesis An Exploration of Big Data and Analytics Software		
Instructor Sakari Kinnunen		**Pages** 53
Supervisor Sakari Kinnunen		

Big data has become a significant part of the modern society and influences all aspects of life ranging from society, economy, to science. Large organizations have implemented big data analytics for management, maintaining production processes, developing products, marketing, avoiding risks, etc. Nevertheless, the degree to which small and medium sized entrepreneurs (SMEs) utilise data or big data analytics is surprisingly limited. This thesis thus sets out to study, among others, the many issues that create barriers that prevents SMEs from implementing big data analytics. The issues included lack of understanding, dominance of domain specialist, cultural barriers and intrinsic conservatism, shortage on in-house data analytic expertise, shortage of useful and affordable consulting, non-transparent software market, lack of intuitive software, concerns on data security & concerns about data protection and data privacy.

This study aimed to describe the state-of-the-art of the Big Data management with a special focus on the methods and tools of (big) data analytics. The theory part discussed big data definition, described five selected data analytics methods and introduced three popular analytics software programs. The practical part of the study explored the features of different Big Data management/ analysis software programs and identified the challenges faced by SMEs when implementing Big Data analytics. Subsequently, a comparison study of these different features was conducted. Based on the findings, this study proposed a recommendation to Centria University of Applied Sciences regarding the most suitable analytics software programs for the university’s corporate partners.

Key words
Big Data, Business Analytics Software, Data analytics, SMEs.

ABSTRACT

CONCEPT DEFINITIONS

CONTENTS

GRAPHS

FIGURES

TABLES

Concept Definitions

Concept	Definition
APIs	Application programming interfaces
AWS EMR	Amazon Web Services Elastic MapReduce
EDI	Electronic data interchange
Field data	Data which are collected outside of traditional workplace
IC	Independent cascade is an information diffusion model where the information flows over the network through cascade
IoT	Internet of thing
IT asset	A piece of software or hardware within an information technology environment
Linguistics	The scientific study of language
LT	Linear threshold
Node	A computer can be used for processing and storing
ODBC	Open Database Connectivity OT – Operational technology. A software or hardware that detect or causes a change
REST	Representational state transfer
RFID	Radio-frequency identification
SMEs	Small and medium sized entrepreneurs
SQL	Structured Query Language
SWIFT	A programming language
Tie	A relationship between nodes
URI	Uniform resource identifier
XML	Extensible Mark-up Language

1 INTRODUCTION

“*Without big data, you are blind and deaf and in the middle of a freeway.*”
- Geoffrey Moore.

The concept of Big Data has attracted attention not only from practitioners but also from management researchers. While the concept creates a big buzz, the mechanism in which big data function remains unclear to many. This thesis, thus, aims at providing a comprehensive review of the concept and its fundamentals. Accordingly, the aims of the thesis are: (i) to introduce Big Data and (ii) to specify aspects to consider in order to choose a suitable software for an organization. The thesis is organized into two parts: theory and practice. Details of the two part are discussed as follows.

The theory section of the thesis will, firstly, present some fundamental knowledge about big data. Secondly, it will discuss the operational and strategic significance of big data analytics. Finally, this chapter will present a brief description of the main research questions and the goals that the thesis attempts to achieve. The practice part introduces a number of challenges that small - medium sized entrepreneurs (SMEs) face in the implementation process of big data. Subsequently, it will describe the data security system and internet of things (IoT) of selected software programs from the theory part. In the final part, based on both the theory and practice parts with features, the conclusion will be presented including software suggestions.

1.1 Background

In the last decade, both the popularity and prevalence of *big data* have been “booming.” It has brought numerous opportunities to various industries, e.g. social networks, manufacturing, healthcare, service, banking, etc. Big data enables organizations to become more operatively efficient and managerially effective. Besides the evident benefits, big data also create challenges to data scientists and computer developers. Such challenges include finding the optimal way to get values and insights from data from trustworthy sources with high security. The goal of this thesis is to provide some fundamental knowledge about big data and the benefits big data can bring to society and organizations. Furthermore, in the practice part, this thesis introduces some data analytics software programs and the functionalities they can offer to organizations. More specifically, this thesis will compare the features of the software programs. This part specifically aims to provide SMEs with some basic guidelines on how to choose a suitable data analytics software programs for their organizations given their specific needs. In addition, this thesis will also propose a recommended software program for Centria University of Applied Sciences.

1.2 Motivation and research questions

In this Industry 4.0 era, data impact all businesses and industries. The ways an organization utilizes data can create or compromise its competitive advantages. Therefore, it is essential to provide a guideline on how to access reliable data analytics and, consequently, help decision-makers make timely and better-informed decisions. As an answer to the complexity of big data, many data scientists and computer developers have been innovating various methods and software programs to analyse data.

Data analytics has strategic significance in operations management. It provides organizations with tools and algorithms that detect failures, analyse and minimise the impacts of risks, forecast sales of a new product, etc. Besides, as the volume of data is becoming increasingly bigger, using only traditional data management and processing methods is no longer sufficient in operations. Inefficient and/or ineffective utilisation of data could result in the unnecessarily high direct and opportunity costs of capital to invest and maintain devices such as hardware. (Hurwitz, Nugent, Halper & Kaufman 2013, 46.) For that reason, the purpose of this study is to provide fundamental knowledge about big data across various sectors of business.

Accordingly, the research questions (RQs) are as follows. There are two overarching research questions that correspond to the two main pillars of this thesis: (i) theory and (ii) practice. The two main research questions are subsequently divided into a total of five sub-questions.

- RQ 1: What is big data and big data analytics?
    - **1.1.**What is the definition of big data and its characteristics?
    - **1.2.**What are the types of big data?
    - **1.3.**What are data analytics?

- RQ 2: What are the tools for big data analytics and their functionalities?

  2.1. What is an analytics software?

  2.2. What are the benefits of data analytics for SMEs?

The first section of this thesis will concentrate on literature review, providing a general understanding of big data – specifically, what its definition, characteristics, and types are. The second part involves comparing the data analytics software, such as Hadoop, Pentaho, the SAS and Oracle. Subsequently, suitable software programs for SMEs will be recommended.

2 LITERATURE REVIEW

This section of the theory part dicusses the definition and characteristics of big data, how data are created and contributed to big data, the type of big data, and data analytics. Big data is a complex and multifaceted phenomenon that encompasses numerous aspects. Thus, this section does not strive to provide a comprehensive understanding of big data; instead, it only aims to provide the fundamental understanding of big data and its basic characteristics.

2.1. Big data definition

Human beings are creating new data every single passing minute. As a result, everyday a large amount of data is contributed to a big data source. For instance, every post that people make on social media is a contributing factor to the ever-increasing big data. In addition, an email to a friend or an organization can also be tracked to generate data about, e.g. the volume of interactions between people or the communication style of a particular individual. Similarly, activities on web browsers can also be tracked to generate targeted advertisements. Consequently, every member of the modern society is an active source of big data. Nevertheless, what exactly is *Big Data*? This part will review a number of definitions that are related to Big Data.

> ‘‘*Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation*.’’
>
> - Gartner IT Glossary, n.d.

According to Interactions with Big Data Analytics (2012) by Microsoft, big data is data that is enabled to be managed and processed by simple methods. Conventionally, data are stored in a database, often in the form of a memory, that can be easily accessed; the values and cleanliness of the data can also be evaluated and determined quickly. Nonetheless, it is impossible to do the same with big data because big dataset is often too large to fit in a memory. Computations involving big data will also take longer as big data cannot be processed with a simple system. For instance, data generated from an image, an audio or human language often require a highly complex algorithms and high computational capability.

Alternatively, processing big data with a simple system often results in unreliable results. In addition, big data might be too large to be saved in one single hard drive; as a result, big data could not be processed in unification. (Fisher, DeLine, Czerwinski & Drucker 2012.)

According to Fisher et al. (2012), big data have three main characteristics: (i) volume, (ii) velocity, and (iii) variable. These three characteristics of big data are collectively called Three V's (V3) (See Figure 1).

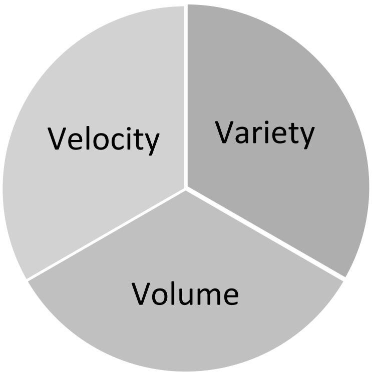


FIGURE 1. Characterizes Big Data by its velocity, variety, and volume – or simply, Three V's

The first V of the Three V's, ***Velocity,*** is the speed of processing data. In other words, data is handled quickly with accurate results. Wal-Mart, a USA discount retail chain, provides a quintessential example of the velocity of big data. Accordingly, Wal-Mart processes more than one million customer transactions every hour with databases calculated over 2.5 petabytes. (The Economist 2010.) Another example of the velocity of big data is the Google search engine - the most popular search engine in the world in 2019. A simple search of the keyword "big data" on Google yields 7,370,000,000 results in 0.7 seconds (Google April 2020).

The second V of the Three V's, ***Variety***, describes the multi-structured nature of big data (i.e. the various types of data). By definition, big data are usually contributed from numerous sources and in disparate formats. For example, structured data (e.g. relational data: database, spreadsheet), unstructured data (e.g. human language text), semi-structured data (e.g. RFID, Extensible Markup Language - XML), and streaming data (e.g. data from machines, sensors, web application, and social media). (Russom 2013.)

The last V of the Three V's, ***Volume,*** refers to the quantity of data. Facebook users upload over 260 billion images – the equivalent of over 20 petabytes of data and about 1 billion photos, 60 terabytes uploaded each week. Google, Amazon, Microsoft, and Facebook all together could hold up at least 1,200 petabytes data between them. Besides, other big companies such as Dropbox, Barracuda, SugarSync, and Instagram also store a large amount of data. (Beaver, Kumar, Li, Sobel & Vajgel 2010.) The data inflation table below will give some examples of the unit in data.

TABLE 1. Data inflation (developed from The Economist 2010, 3)

Unit	**Size**	**What it means**
Bit (b)	1 or 0	Short for ‘‘binary digit’’, after the binary code (1 or 0) computers used to store and process data
Byte (B)	8 bits	Enough information to create a number or an English letter in computer code. It is the basic unit of computing
Kilobyte (KB)	1000K; 2^{10} bytes	From ‘‘thousand’’ in Greek. One page of typed text is 2KB.
Megabyte (MB)	1000 KB; 2^{20} bytes	From “large” in Greek. The complete works of Shakespeare total 5 MB.
Gigabyte (GB)	1000 MB; 2^{30} bytes	From ‘giant’ in Greek. A two-hour film can be compressed into 1-2GB.
Terabyte (TB)	1000 GB; 2^{40} bytes	From ‘monster’ in Greek. All the catalogued books in America’s Library of Congress total 5TB.
Petabyte (PB)	1000 TB; 2^{50} bytes	All letters delivered by American’s postal service in 2010 was amount to around 5PB. Google processed around 1PB every hour in 2010.
Exabyte (EB)	1000 PB; 2^{60} bytes	Equivalent to 10 billion copies of the Economist.
Zettabyte (ZB)	1000 EB; 2^{70} bytes	The total amount of information existence in 2010 was forecasted around 1.2ZB and 2018 was approximately 33ZB.
Yottabyte (YB)	1000 ZB; 2^{80} bytes	Currently too big to image.
The prefixes are set by International Group, the International Bureau of Weights and Measures.		

In addition, there is also an extended version of Three V’s called ***the Five V’s***. The Five V’s has two additional characteristics of Big Data: **Value** and **Veracity**. These two characteristics are discussed as follows.

Value. Oracle claimed that data always have internal value; however, data can be used only after that value is discovered. Many big companies see big data as their capital, which constitutes a large part of the value. Value is created when data are analysed and contributed to the development of new products and/or services. (Oracle 2013.)

Veracity. Veracity describes the trustworthiness of big data. More specifically, veracity describes the percentage of the analytics or prediction that can be trusted. The output data is less trustworthy if it comes from an unreliable source, or if it is based on too little input data. Nowadays, the advancement of technology allows users to have easy access to a higher volume of data. As a result of technological advancement, the costs of data acquisition and accessibility also decrease significantly over the years. Owing to the large amount of data, data analyses are getting more accurate and reliable. (Oracle 2013.) For example, the Internet allows consumers to have easy access to information about a particular product or service. This consequently allow consumers to research about the product or service before making a purchase. The more reviews there are, the more accurate the perception a consumer will have about the product or service will be.

2.1.2 Type of Big Data

Big data has three types of data: structured, unstructured and semi-unstructured data. The three types of data will be discussed in more detail in the following paragraphs.

Structured data are categorized and formatted data in a relational database, a file, or a spreadsheet in tabular form. For example, a customer's list, inventory or purchasing list, etc. In this case, all information has been organized and formatted; thus, it is easy to search or analyse. According to many data scientists, structured data constitute for around 20% of all data stored. (Hurwitz et al. 2013, 26.)

Unstructured data is heterogeneous data, which do not have a particular format, such as human language, text, photographs, or video. Unstructured data expand faster than structured data. IDC study (2018) claims that by 2025, the amount of data is predicted to be increased tenfold compared to the amount in 2016. Accordingly, by 2025, the amount of data could reach 175ZB. According to Reinsel et al. (2018), unstructured data constitute for approximately 80% of all data stored.

Semi-structured data is in between structured data and unstructured data, which does not follow any standards. Semi-structured data can be Extensible Markup Language – XML and Radio-frequency Identification (RFID), EDI or SWIFT. (Hurwitz et al. 2013, 30.)

2.2. Data analytics

A big volume of data is created every day. These data are needed to be processed and analysed to generate the value and create benefits. Thus, big data analytics attracts great attention from data experts and organizations. There are three main purposes of data analytics:

- Improve decision making.
- Minimize risk.
- Bring out valuable insights (Tumer 2016.)

This section will separate data analytics into five categories as follows: (i) text analytics, (ii) audio analytics, (iii) video analytics, (iv) social analytics and (v) predictive analytics.

2.2.1. Text Analytics (text mining)

Emails, social media, news, surveys, online forums are examples of textual data; and most of them are unstructured data. Text analytics is a term to transform human language text into high-value information. For instance, if a user search for information about a particular item on a search engine (e.g. Google), the information about this search could be analysed and used for targeted advertisements. Subsequently, when the same user browses social media such as Facebook or Instagram, he/she will see some advertisements of the same item or related items.

Processing of text analytics includes Natural Language Process (NLP), Information Extraction (IE), and Text summarization. Natural Language Process describes the use of computational methods to process unstructured text (spoken or written form) which take a role as a mode of communication commonly used by humans. (Assal, Seng, Kurfess, Schwarz & Pohl 2011.) Information Extraction systems transform unstructured text input to structured data specified under different criteria or catalogues. With significant and detailed description of the linguistic content, IE can be useful to contribute more structured input into databases. Information Extraction (IE) refers to the use of computational methods to determine important information in documents and transform this information into representation, which is suitable for computer-based storage, processing and retrieval (See Figure 2). (Wimalasuriya & Dua 2010.)

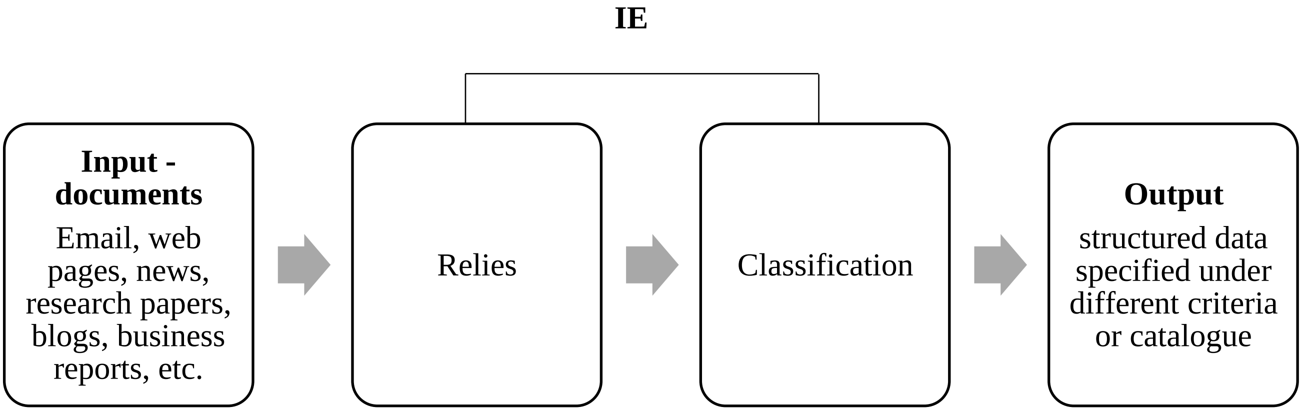

IE
Input - documents
Email, web pages, news, research papers, blogs, business reports, etc.
Relies
Classification
Output
structured data specified under different criteria or catalogue

FIGURE 2. Information Extraction (IE) system

IE includes the following sub-tasks:

- Entity Recognition. ER searches for names in the text and separates them in specific catalogues or folders.
- Relation Extraction. RE detects and extracts the relationship between nodes.
- Coreference Resolution.

Text summarization summarizes one or many documents in order to condense important information in the original text(s). Text summarization applications can be used widely in, e.g. articles, advertisements, news, reports, etc. The two main text summarization methods are: (i) the extractive approach and (ii) the abstractive approach. The extractive approach creates a summary by using important texts (typically sentences) from the original documents. Important texts and sentences are defined by their frequency and location in the texts. Hence, the systems do not need to '*understand'* the text. Instead, the systems only summarize the texts through skimming, scanning and extracting the key information. (Gandomi & Haider 2014.) In contrast, an abstractive approach needs to 'understand' the text to extract semantic information, i.e. the idea of the text. The summary might not contain the sentences from original texts. In this approach, the Natural Language Processing (NLP) is used to parse and create the summary. (Gandomi & Haider 2014.) As a result, an abstractive approach creates a more understandable summary than extractive approach. However, an extractive approach can be used more easily, especially big data with large amounts of data. Table 2 shows the distinction between extractive approach and abstractive approach.

TABLE 2. Extractive approach versus Abstractive approach

Extractive approach	**Abstractive approach**
• Systems do not need to understand the text • The summary contains text units from the original text (sentences) • Easier to be used	• Systems need to understand the text • Using NLP • Summary might not contain text units from the text (sentences) • Summary is more understandable

2.2.2 Audio analytics

Human spoken language is identified and stored as audio data. Audio analytics or speech analytics are used to analyse and get information from audio data. This application tends to be popular. Audio analytics is mostly used in the customer call-centre or healthcare. In general, call-centres and healthcare use this application in order to improve customer service, understand customer behaviour, etc. Below, two popular technologies are introduced.

Large-vocabulary Continuous Speech Recognition recognizes the sounds of the words and searches in their dictionary/database. In case they cannot find the same word in the dictionary, it will find the most similar one. Subsequently, the system will find the file or information about that word. (Gandomi & Haider 2014.)

The phonetic-based system is used to distinguish units of similar sounds, such as *look* and *took* (Gandomi & Haider 2014). The applications of such system can be observed through voice recognition technologies, such as Google's voice search tool, Siri of Apple, dictation language on texting, etc. For instance, *reign* and *rain* have similar pronunciations and can be often mistaken for each other. Nevertheless, when a user uses Google's voice search tool and says, e.g. "the king reigns the country" and "it rains", Google is still able to recognise the word correctly as it uses audio analytics to analyses and chooses the option with a higher possibility based on the context of the whole sentence.

2.2.3 Video analytics

Video analytics is used to get information out of videos and mostly used in automated security and surveillance systems. Traditionally, security surveillance often involves personnel at the scene to ensure everything is under control for safety reasons. (Valentín, Serrano, García, Andrade, Palacios-Alonso & Sucar 2017.) For instance, security guards at railway stations often need to patrol around the stations periodically. However, Closed- Circuit Television CCTV provides a better solution for such purpose. Accordingly, CCTV sends information to the control centre – where personnel inspects. If something happens, the controller is able to act quickly, such as a call for help, turn on the alarm, etc. Moreover, the video can be saved for inspection reasons. In some advanced video analytics systems, it is possible to analyse complicated input such as human behaviours or movements. Some advanced systems are even able to predict emotions based on behaviours. (Gandomi & Haider 2014.)

The data analytics enabled by CCTV from retail stores can also be used for business insights. The organizations can study from CCTV to get to know consumer behaviours. Advanced video analytics systems (e.g. facial recognition) can collect information about customers, such as gender, age, detect customers movements, period time they spend in the store, etc. (Gandomi & Haider 2014.) The output data can support a company's strategies or decisions regarding marketing, promotion, up-selling, price, etc. According to Video analytics – choices of architecture by Mr. Kiron Kunte from Norik consult, server-based and edge-based are two most used architecture in video analytics. In Server-based architecture implementation, the cameras capture the video and send it to a dedicated analytic server, where the video is analysed, and the issues/problems are identified. Subsequently, the server will send analysis outcomes and alert signals if necessary. The video analytics intelligence is located on the server. (Kunte 2012.) Unlike the server-based architecture, in the Edge-based architecture, the cameras have two functions: capturing and analysing the video. For that reason, the cameras or encoders have to be powerful enough in order to concurrently record and process video analytics. (Kunte 2012.). Table 3 illustrates the differences between a server-based architecture and an edge-based architecture.

TABLE 3. Server-based architecture versus Edge-based architecture

	Server-based architecture	**Edge-based architecture**
Pros	• Can be used with most cameras. • Cameras options are various. • Better performance. • Have more analytics applications options. • Powerful processing. • Upgrading system easily. • Be set up and used easily. • Cost effective	• Reduced bandwidth usage. • Preventing send all video data to server.
Cons	• Large bandwidths usage.	• Unable to save all videos for a certain period of time/ motion-detected events. • Costly camera, devices. • Might lead to limited performance or basic analytics application. • Lower processing power.

2.2.4 Social media analytics

Social media analytics involves analyses of data acquired from social networks, e.g. blogs, social news, networking sites, etc. The key characteristic of social media analytics is its data-centric nature. Due to the growing popularity of social media, social media analytics is becoming a powerful tool for corporate decision-makers. Social media analytics can be categorized into five main types: (i) content-based analytics, (ii) structured-based analytics, (iii) community detection, (iv) social influence analytics, and (v) link prediction.

Content-based analytics analyses information by converting it from content. Structured and unstructured data, e.g. emails, videos, texts, databases, are collected and transformed into analysed information (Bari, Chaouchi & Jung 2017,41.) Content analytics software programs use human language, predictive analytics, trend analysis, contextual discovery, benchmarks, key performance indicators, etc. to analyse data to get meaningful insights of for strategic decision-making processes. Content-based analytics can help detect consumer preference and behaviour and thus provide decision-makers with useful insights for e.g. new product development, marketing and advertising, etc. (Zhu, Foyle, Gagné, Gupta, Magdalen, Mundi, Nasukawa, Paulis, Singer & Triska 2014.)

Structured-based analytics (i.e. social network analytics - SNA) plots and measures the network of relationships and interactions between different information or knowledge entities, e.g. between people, groups, organizations, websites, computers, animals, and nations. The SNA structure is created by nodes and edges (i.e. ties, links), representing humans or groups and relationships or flows. The SNA supports both a visual and mathematical analysis of relationships, which is shown by highly complex, graph-based networks. (Jamali & Abolhassani 2016.)

Community detection can be seen as data mining on graphs. Community detection can be used to detect similar groups of nodes in the network and collect useful information from these connections. Many networks have been found to contain dissimilar mass of vertices or nodes in different groups. Within these groups, there are many ties between nodes, but there are fewer links between groups. In this case, community discovery algorithms can indicate the connections between groups by maximizing the number of ties insides the groups while minimizing the number of ties between the groups (i.e. modularity in the network). The goal of community detection is to recognize the connections within groups and the distinction between groups. Consequently, it could help a company or brand understand, e.g. the differences across different consumer groups in terms of demands, preferences, and perceptions towards its product or service offerings. Determining accurately the target groups of consumers leads to a more effective marketing and advertising strategy. (Coscia, Giannotti & Pedreschi 2012.) For example, Instagram can recommend sponsored advertisements through its "*base on posts you have shared"*-algorithm or based on user activities, e.g. whom he/she follows on Instagram. For instance, if the user is interested in outdoor activities and follows Instagram accounts that are related to outdoors activities, Instagram will recommend advertisements on related topics to the users.

Social influence analysis is used to evaluate the influence of an individual and that individual's connections in a social network by models, e.g. independent cascade – IC, linear threshold – LT, and algorithms, e.g. influence maximization, influence minimization, flow of influence, individual influence. The information can give significant insights for many decision-making areas, such as viral marketing, online recommendation, e-business, etc. (Li, Zhang & Huang 2017.)

Link prediction predicts new links among the existing links of the nodes by similar structural metrics, e.g. the number of connections. Accordingly, pairs of entities with higher similarity or closeness are more likely to be connected to understand and predict the effectiveness of the network. Link prediction can be used in many applications. (Zhou et al. 2018.) For instance, Facebook and Instagram recommend new connections to their users (i.e. *"People you may know"*) based on the number of mutual connections a user has with the potential new connections. Similarly, Netflix recommends its users with TV-shows and movies based on data about the users' past activities. Accordingly, Netflix shows how similar the recommendations are to those past activities through *"percentage matching."* In an essence, Netflix analyses the link across different movies/TV-shows and utilises these links to improve its retention rate. In addition, inferences from link prediction can be used to detect and uncover criminal and terrorist connections.

2.2.5 Predictive analytics

Predictive analytics essentially acts a data-driven fortune teller by giving predictions on what might happen based on historical and current facts/data. Thus, predictive analytics provides a great support for organizations in various operational and strategic aspects, e.g. demand forecasts, profit maximization, risk management, operations, simulations, etc. In practice, many organizations spend a large portion of their budget on managing risks in order to reduce negative impacts on their organizations. Risk management is an essential part of operations management that will ultimately help organizations to retain their competitive advantages in their respective markets. Therefore, predictive analytics is relevant in most, if not all, industries and markets. For example, predictive analytics can predict failures of engines based on stream data from sensors. The results from this type of predictive analytics would allow companies to schedule timely maintenance in-advance and consequently, decrease operational costs (e.g. costs of breakdowns, costs of lost service, etc.), decrease maintenance lead-time (i.e. time to wait for maintenance to arrive) and improve their service levels. Among the most-discussed topics in predictive analytics are data mining and machine learning. (Bari et al. 2017.)

Data mining. Data bring benefits only if one knows how to exploit their hidden values. Valuable data are often not visible but hidden and just like precious metals, organizations have to dig to find them, hence the term "*data mining*". Data mining aims to explore secret patterns by using mining tools through machine learning. Mining tools can be sophisticated algorithms, e.g. classification trees, logistic regression, neutral network, clustering techniques like K-nearest neighbours. (Hurwitz et al. 2013, 145.)

Machine learning. The purpose of machine learning to explore knowledge from data and provide significant insights for decision-making processes. The machine learning algorithm is able to conduct predictive analytics. For example, customer behaviour can be studied through analysing their purchasing history, history of searched items, and even what they say or what they are interested in on social media. As a result, the organization can provide different target groups with different offers/ products based on micro-segments of similar customers. (Bari et al. 2017, 42.) A predictive analytics process is a combination of data mining and business knowledge to determine hidden values (See Figure 3).

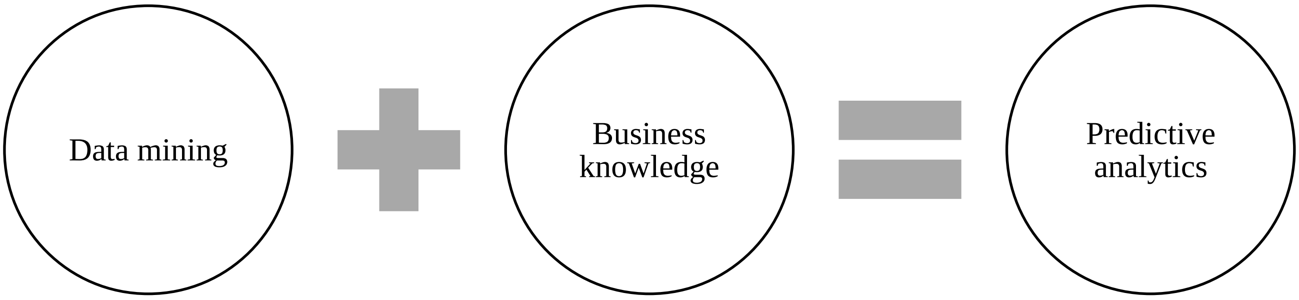


FIGURE 3. Predictive analytics (developed from Bari et al. 2017, 20)

2.3 Data Visualisation

Big data contains a large volume of information. Such information is crucial for operational and strategic decision-making regarding e.g. marketing activities, supply chain management, etc. Nevertheless, as the volume of both data and information is enormous, it is highly challenging to see the truly important pieces of information. Data visualisation offers a solution to such challenges. In an essence, data visualisation helps communicate the insights generated from big data analytics to decision makers in a more reader-friendly manner than complex algorithms.

Data visualisation has a significant role in communicating insights from data through images, diagrams, and animation. For instance, it can illustrate a trend through time series, comparing and ranking, correlation, etc. In the domain of data visualisation, Tableau is among the most popular tools. Nonetheless, most analytics software programs provide data visualisation as an integrated functionality. Besides Tableau, there are many other options for data visualisation, e.g. Pentaho, SAS, etc. (Tableau software.)

In addition, big data and data analytics are usually linked together by interactivity. Data visualisation can illustrate the relationships between different values in the same category or the relationship among specific areas. There are many options for data visualisation, such as charts, scatter plots (i.e. scatter diagram), box plots, etc. When doing data visualisation, it is necessary to choose the right type of visualisation for the right types of data. The right type of charts will improve the effectiveness of data visualisation; conversely, inappropriate choices of charts can be detrimental to data visualisation and hinder its effectiveness. (Tableau software.) Thus, knowing what types of charts work best for what type of data and analyses is the basis of effective data visualisation.

3 ANALYTICS SOFTWARE

The organisations are not able to extract valuable data and insights from large amounts of data without analytics software programs. That is a must-have tool for any business to transform data into useful information. As a result, data science has become a competitive segment in the current economy. Technology companies have been continuously innovating and developing products and solutions to meet the demands from the market. A quick search on the Google search engine with the keyword *"big data analytics software*" would yield numerous software programs. This thesis will review four software programs: (i) Hadoop, (ii) Pentaho, (iii) the SAS and (iv) Oracle.

3.1. Hadoop

Data are traditionally stored in hardware, which causes two big problems. Firstly, there needs to be hardware with sufficiently large capacity, which is impractical and costly. Furthermore, it is also risky because if an error occurs, the data could be completely lost. Secondly, an analysis is often a result of a combination of data from various sources. This traditional way could take time and the analysis might not be accurate. Consequently, Hadoop was developed to tackle these problems. (White 2012, 3.)

Hadoop is an open-source software structure for storing data with numerous distributed systems to process big data on computer clusters. In other words, Hadoop stores, organizes and manages big data in the easiest way for users. Hadoop helps computer scientists process big data more easily. It also fixes the problems, or at least controls the failures, occurred while processing large amounts of data. In addition, Hadoop provides an inexpensive solution for organizations. Hadoop's characteristics are as follows.

- Hadoop is a framework, which allows developers to develop distributed application.
- Hadoop is written in Java. Thanks to streaming data, Hadoop allows developers to develop distributed application in Java and other languages such as C++, Python and Pearl.
- Hadoop provides a storage method with distributed data on many nodes.

- Hadoop can be run on the GNU/Linux platform. Windows is also a supported platform; thus, Hadoop can be used on Windows. (White 2012.)

The seven features of Hadoop, (i) Hadoop Distributed File System, (ii) Hadoop MapReduce, (iii) Hadoop YARN, (iv) Hadoop Pig, (v) Hive, (vi) ZooKeeper, and (vii) Sqoo, are described in the following section.

3.1.1. Hadoop Distributed File System (HDFS)

Hadoop Distributed File system allows users to distribute the storage of big data through cluster computers, running on commodity hardware. All hard drives function as a big file system and maintain redundant copies. Besides, HDFS support optimising data bandwidth usage between nodes. The main functionalities of HDFS are described below.

- Back up data. In case something happens to the hard drive, HDFS can detect the faults and automatically recover quickly all data.

- Streaming data access is needed when applications run on HDFS.

- Support large data files. Hadoop clusters can store very large data, which could be petabytes in size. Dividing the data set into a small number of large files could be more optimal. It can support reducing data access time and simplifying file management.

- Commodity hardware. HDFS is made to run on clusters of commodity hardware. For large clusters, there is a higher chance of failure of nodes across the cluster, but HDFS continues working without notification interruption to the user.

- HDFS applications need a write-once, read-many-time model. Most files are changed by appending the content to the end of files rather than overwriting existing data. This preserves data consistency and allows data access with high throughput. (White 2012, 45 - 46.)

Nevertheless, HDFS is not suitable in circumstances that have, e.g., a low/latency data access, lots of small files, multiple writers, arbitrary file modification.

3.1.2. Hadoop MapReduce

Hadoop MapReduce is a software framework that allows processing large data in-parallel across clusters. There are two independent parts: Mapper and Reducer. The Map component distributes the programming problems or tasks and transforms data into parallel data across clusters (large data system) with an efficient method. After the distribution is completed, the Reducer aggregates all the data together to provide a result (See Figure 4, 5, and 6). (White 2012, 525.)

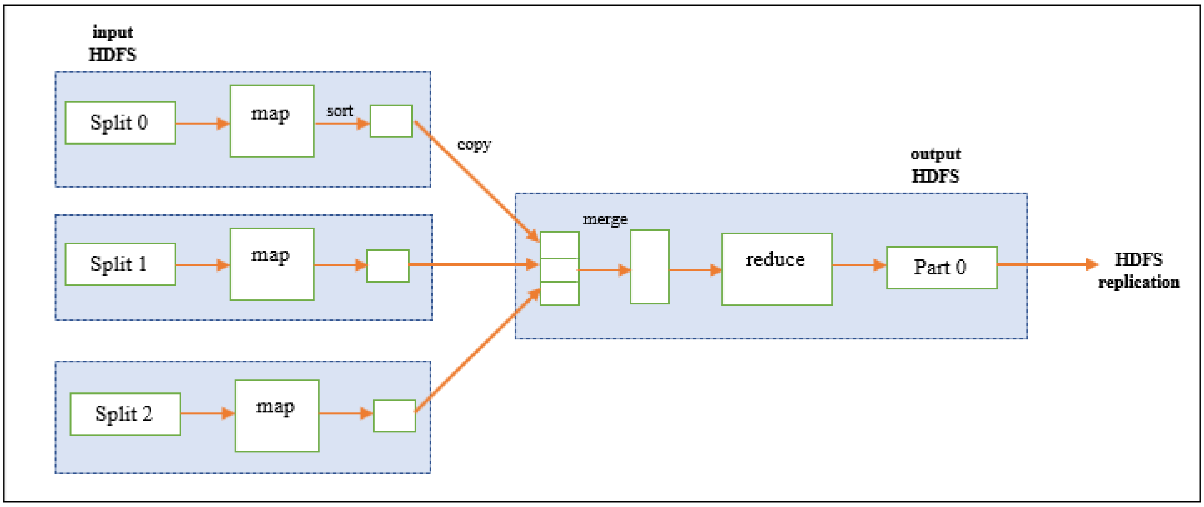


FIGURE 4. MapReduce data flow with a single reduced task (developed from White 2012, 33)

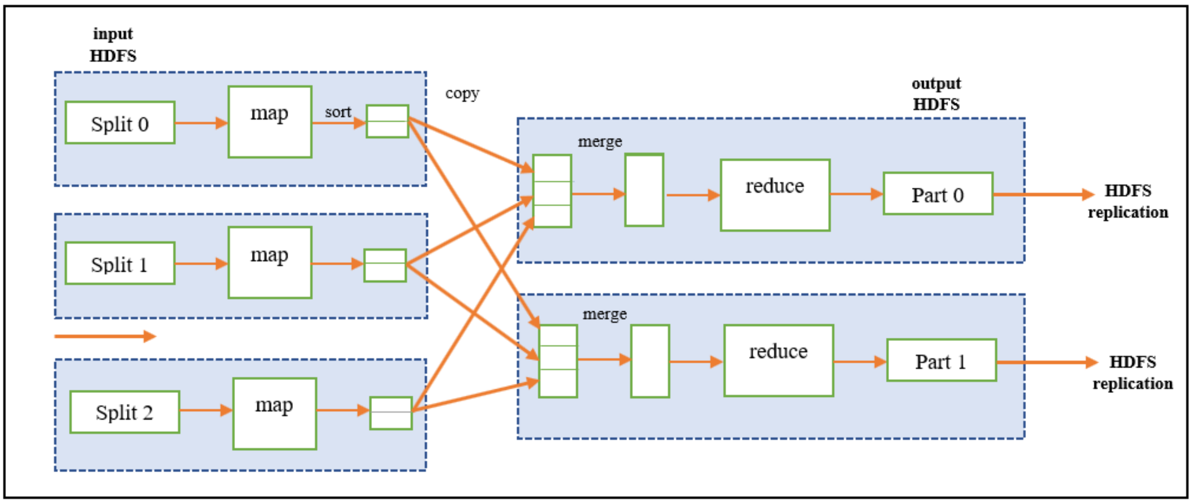


FIGURE 5. MapReduce data flow with multiple reduce tasks (developed from White 2012, 34)

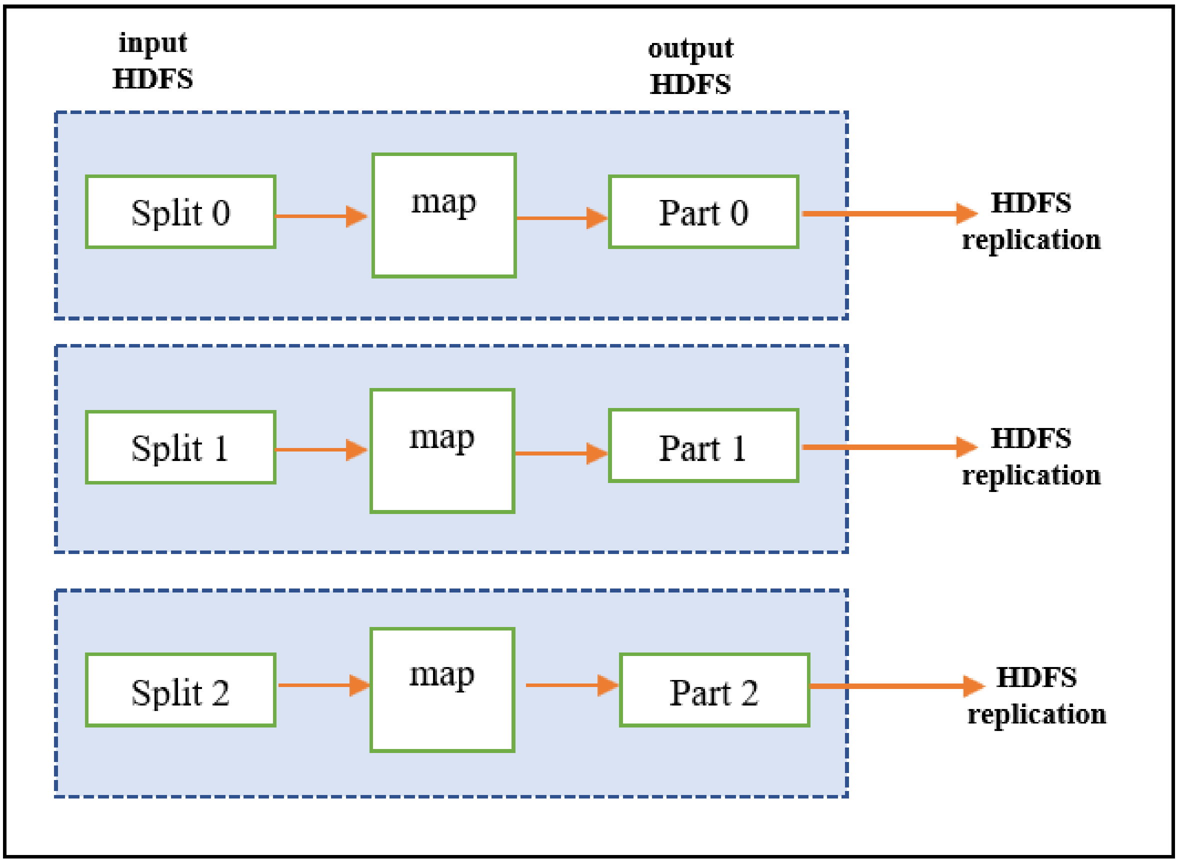


FIGURE 6. MapReduce data flow with no reduced task (developed from White 2012, 35)

3.1.3. Hadoop YARN

Hadoop YARN stands for Yet Another Resource Negotiator**.** Hadoop YARN aims to manage the resource on a cluster, which decides (i) what gets to run tasks, (ii) when the nodes are available for extra work, and (iv) which nodes are unavailable. In addition, YARN is an application master to manage the lifecycle of applications running on the clusters. (White 2012, 195.)

3.1.4. Hadoop Pig

Hadoop Pig provides a high-level scripting language that sits on top of MapReduce. Hadoop Pig is used for discovery of large datasets. Pig transforms the scripts to run on MapReduce, which goes through YARN and HDFS to process the data and give a result. With a very high-level programming API, users can write a simple script, but it looks like SQL (Structured Query Language) in some cases. This script allows the program to connect queries and get complex results without writing Python or Java code. Pig provides a high-level data-flow language and execution framework for parallel computation. Pig is created from two parts:

- Pig Latin is the language used to convey data flows.

- Two execution environments to run Pig Latin programs. Local execution in a single JVM and distributed execution on a Hadoop cluster. (White 2012, 365 – 366.)

3.1.5. Hive

Hive is a framework for data storage for Hadoop. Hive is similar to Pig, but simultaneously, it also resembles an SQL database. Hive is designed to make it possible for analysts with strong SQL skills to run queries on the large amount of data. Hive can be connected through hive client and ODBC (Open Database Connectivity).

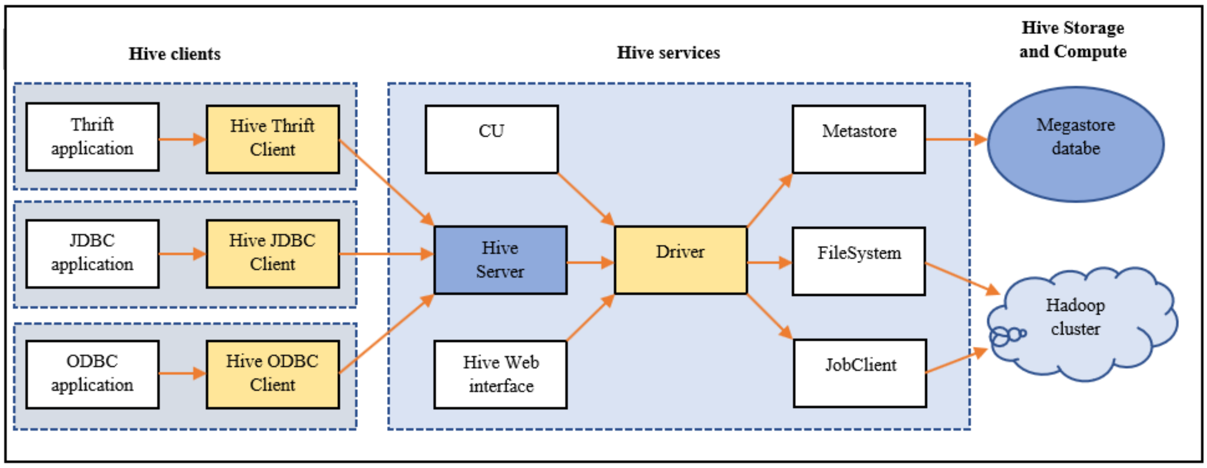


FIGURE 7. Hive architecture (developed from White 2012, 418)

One of the limits of Relational Database Systems is that it does not allow systems to host large, sparsely populated tables on clusters made from commodity hardware. From that limitation, HBase was invented to solve the problems by giving proper problem space. HBase is not relational and does not support SQL. (White 2012, 457.)

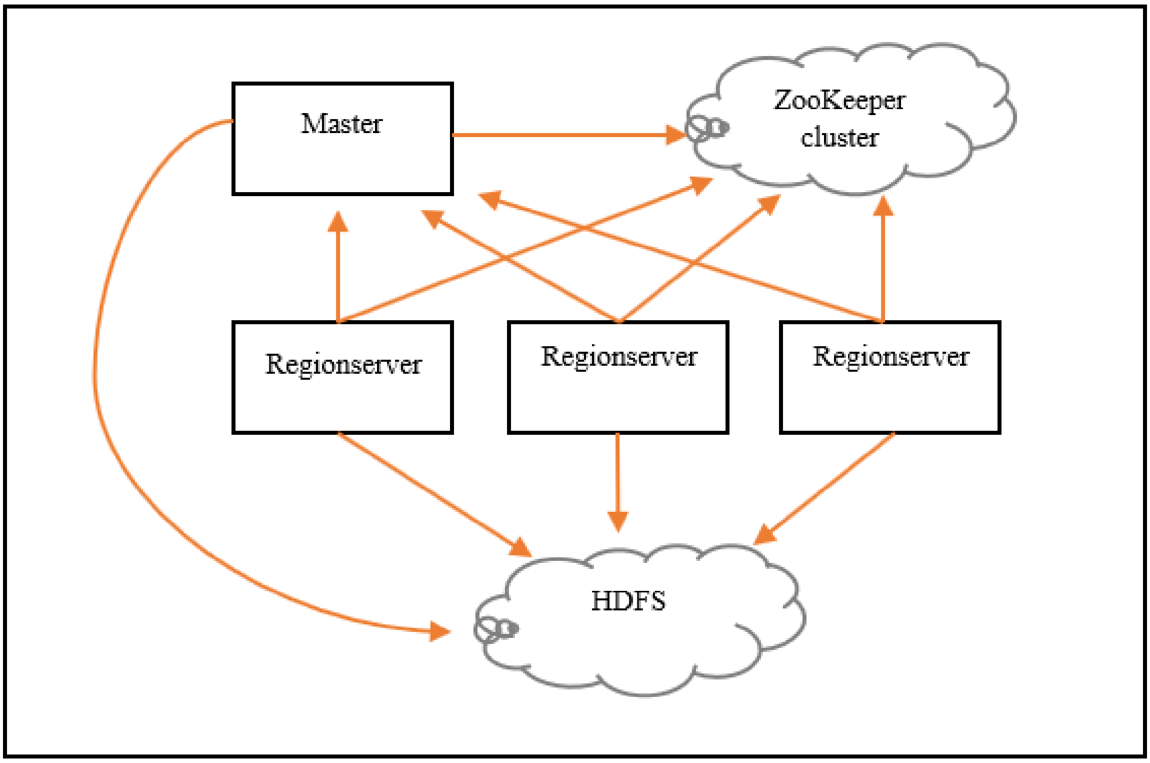


FIGURE 8. HBase cluster members (developed from White 2012, 457)

3.1.6. ZooKeeper

Zookeeper is a high-performance Hadoop’s coordination service for distributed applications. Zookeeper provides a set of tools to build distributed applications, which can deal with partial failures. (White., 2012, p. 487). In a partial failure, the users are not aware of the failure as there is no warning to the users. For example, if the network fails while messages are being sent between two nodes, the messages might not be sent. Nevertheless, in such circumstances, the users are not notified of the failure. Unless the sender asks the receiver about the message, the sender will remain unaware of the failure. Therefore, the ability to deal with partial failures are crucial for an application to function seamlessly. (White 2012, 487.)

The basic characteristics of ZooKeeper are described as follows.

- Simple: simple operations, and some extra abstractions such as ordering and notifications.

- Expressive: can be used to build a large class of coordination data structures and protocols.

- Highly available*:* ZooKeeper was created to be highly available so applications can rely on it.

- A library. ZooKeeper contributes an open source. From time to time, users can build and improve the libraries.

- Facilitates loosely coupled interactions. Participants can discover and interact with each other without knowing them. (White 2012, 488.)

3.1.7. Sqoo

Hadoop has the ability to work with many forms of data. Nonetheless, MapReduce programs need external APIs to get to that data in storage repositories outside of HDFS – data storage piece of Hadoop. Normally, valuable data in an organization is stored in RDBMS (Relational Database Systems). Sqoop is an open-source tool that can help users extract data from RDBMS into Hadoop for processing. When the final pipe analyse is available, Sqoop can export these results back to the database for consumption by other clients. (White 2012, 525.)

3.2. Pentaho

Data integration (DI) and Business intelligence (BI) are two applications of Pentaho. Pentaho provides consolidated, open, fast, high performance Business Analytics. Pentaho business analytics has the following functionalities: data mining, predictive algorithms, analytics modelling workbench, and visualisation tools. The following parts describe the following three aspects of Pentaho: (i) Pentaho Data Integration, (ii) Pentaho and Machine Learning Orchestration, and (iii) Pentaho Business Analytics Platform.

3.2.1. Pentaho Data Integration (PDI) – an enterprise class and graphical ETL tool

First of all, Pentaho data integration has the ability to send analytics data to clients quickly alongside with the visual tools, which helps decrease time and complication and increase productivity. Normally, the processes are written by coding languages, e.g. Java or Python, or by SQL. Nonetheless, with the Pentaho's code-free data transformation design, the organization does not need those programming languages to write the process. As a result, this could boost productivity by 15 times compared to hand-coding. In addition, users can benefit from the flexibility and insulation from change by Pentaho's adaptive big data layer, which allows the users access to popular big data stores, for example:

- Spark and Hadoop distributions – Cloudera, Hortonworks, MapR, Amazon Web Services Elastic MapReduce (AWS EMR), Google Cloud and Microsoft Azure HDInsight.
- Object stores, such as Hitachi Content Platform.
- NoSQL databases such as MongoDB, Cassandra, and HBase.
- Analytic databases: Redshift, Snowflake, Vertica, Greenplum, Teradata, SAP HANA, Amazon Redshift, Google Big Query, Microsoft Azure SQL Data Warehouse (DW).

- Business application: SAP, Salesforce, Google Analytics.

- Files: XML, JSON, Microsoft Excel, CSV, txt, Avro, Parquet, ORC, unstructured files with metadata, including audio, video and visual files. (Hitachi 2019d.)

Pentaho simplifies and speeds up the process of integrating existing databases with new sources of data. Pentaho Data Integration’s graphical designer includes:

- Making the creation of a data analytics pipeline less complicated by drag-and-drop tool.

- Easy to access to analytics from any processing step to spot check data.

- Composition ability to integrate transformations.

- Ability to access large library, data storages. (Hitachi 2019d.)

Furthermore, Pentaho Data Integration can bring many benefits to customer. The benefits include the following:

- User access to popular data stores.

- Improved performance of data extraction, loading and delivery processes.

- Ability to integrate existing data with new sources.

- Shorter processing time and reduced complexity of integrating big data sources.

- Data profiling and data quality.

- Effective administration and management. (Hitachi 2019d.)

3.2.2. Pentaho and Machine Learning Orchestration

The Pentaho application has the ability to streamline data mining workflow. Thus, it supports data scientists, engineers and analysts in the workflow process. Pentaho maximizes limited data science resources and puts predictive models to work on big data faster, regardless of the use case, industry or language, and whether models are built in R, Python, Scala or Weka. (Hitachi 2019c.) The four key steps of machine learning workflow are illustrated in Figure 9 and described in the following section.

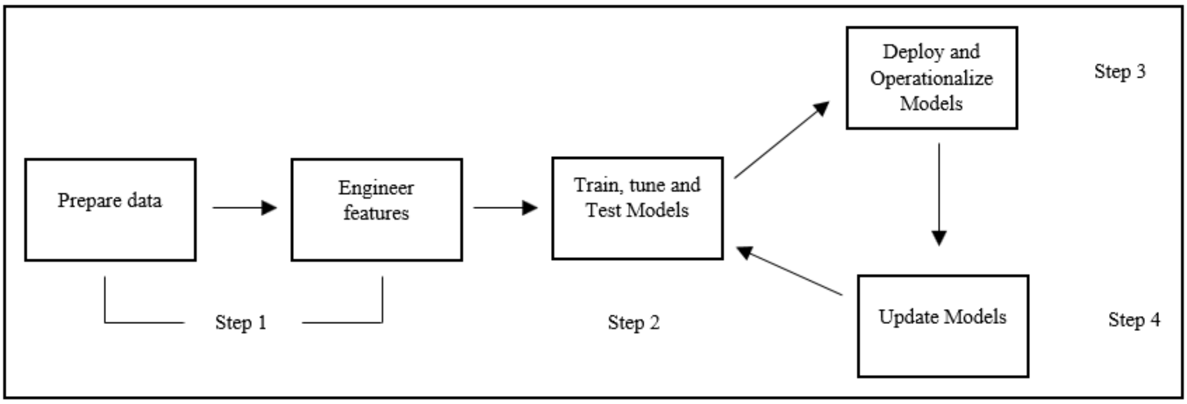


FIGURE 9. Pentaho addresses the four most important stages in the data science workflow (developed from Hitachi 2019c, 1)

Step 1: Prepare Data

As mentioned above, existing data sources, for example, enterprise resource planning (ERP) or customer resource management (CRM), and new data sources, such as sensors and social media, are integrated. Drag-and-drop tools make data analytics faster. (Hitachi 2019c.)

Step 2: Train, Tune and Test Models

Data scientists can work with a seamless and efficient process of train, tune, build and test models thanks to Integration of analytics languages, e.g. R and Python, and machine learning and deep learning libraries such as Tensorflow, Keras, Weka and Spark. Furthermore, popular Integrated Development Environments (IDEs), e.g., Jupyter Notebooks, also support the process. The data programmer can prepare data and create new features or functions and then apply data on the second stage. (Hitachi 2019c.)

Step 3: Deploy and Operationalise Models

Pentaho allows the users to implant models developed by a data scientist as an achievement stage in the operational workflow. Existing data and features engineering efforts can be used in order to reducing development time. With embeddable APIs, organisations can also include the full power of Pentaho within existing applications. The developed models can be used in production environments and transform business. (Hitachi 2019c.)

Step 4: Update Models Regularly

With Pentaho, data engineers and scientists can improve the existing models with new data sets or functions using achievement stages for R, Python, Spark MLlib and Weka. Updating models and achieving the existing ones can be done automatically by prebuilt workflows. (Hitachi 2019c.)

3.2.3. Pentaho business analytics platform

The Pentaho business analytics platform provides many visualisation analytics tools that allow users to create reports, data analysis, data visualisation and dashboards. In addition, thanks to the platform's user-friendliness, users can create analytics by themselves without support from IT engineers or developers. Pentaho Business Analytics provides:

- Ad Hoc Analysis and Visualization.
- Flexible dashboards.
- User-Driven Reporting.
- Mobile Business Analytics. (Hitachi 2018.)

The Pentaho open-source platform makes the process of preparing, integrating and visualising data simple. Thus, anyone in the organization can analyse, visualise, explore, report, predict and transform data into insights or value easily. (Hitachi 2018.)

3.3 SAS

SAS is among the most popular software programs in the contemporary analytics solutions/ software market. According to SAS homepage, the SAS platform provides more than two hundred solutions and products for improving business performances, e.g. advanced analytics, customer intelligence, analytics platform, visual analytics, etc. In addition, the SAS platform can be applied to various industrial segments, ranging from baking, communication, government, healthcare, high technology manufacturing, insurance, media, oil & gas, higher education, to travel & transportation etc. In short, the SAS software program is suitable for various types of businesses. (SAS 2019.)

Moreover, the SAS software program provides many solution packages for different customer groups, such as students & professors, developers, small & medium size businesses, etc. Hence, depending on customer sectors and purpose, the customers or the organisations can choose a suitable software package with a reasonable budget. Last but not least, SAS also provides training programming on the website, allowing all users to access ca. 524 short tutorial videos. The videos offer comprehensive instructions on how to process the data to get results, from putting input to get results. (SAS 2019.) The following part will discuss one integrated part of the SAS platform, the SAS Viya.

SAS Viya is an open analytics platform and an integrated part of the SAS platform. SAS Viya is able to process all kinds of data regardless of their volume, speed, programming language. As a result, SAS Viya allows users to get valuable insights, such as consumer behaviour and preference, operational risks, equipment failure forecasts, etc. SAS Viya is an analytics engine that is able to generate valuable insights in an efficient, accurate and trustworthy manner. There are some benefits SAS Viya can bring to customers:

- Ability to process any type of data and data analytics, such as machine learning, data mining, deep learning, statistics, and artificial intelligence AI.

- Ability to support not only programming in SAS but also other languages, such as Python, R, Java and Lua thanks to a standardised code base.

- Ability to deploy seamlessly and smoothly through all foundations or application systems, thanks to cloud-enabled and in- memory analytics. (SAS 2019b.)

With the aforementioned capabilities, AS Viya is also able to offer the following benefits to customers:

- All teams are given equal power to discover and analyse the same data whenever and however they want.
- As mentioned above, SAS Viya provides a standardised code base. As a result, users can choose that code base or others. This is extremely convenient and easy for users, as they may be more familiar to a particular coding language other than the default one.

- In addition, users can compare different analytics models to find the favourite performance.

- The users can use all analytical tools, such as preparation, interactive exploration, statistics, machine learning, AI, etc. to develop activities. This would result in better performance and higher productivity. Besides, analytical tasks can be done in the same workspace.

- Solve analytics problems. Literary SAS viya is able to solve all analytic problems regardless of their size or complexity. (SAS 2019b.)

In the next page, there is the table of various example options that the SAS can provide to customers depending on the industry and purpose.

TABLE 4. Example products from SAS (developed from SAS 2019b)

SAS' Products	
- SAS Visual Analytics. - SAS Data Preparation. - SAS Visual Statistics. - SAS Visual Data Mining and Machine. - SAS Visual Forecasting. - SAS Optimization. - SAS Econometrics. - SAS Visual Investigator.	- SAS Detection and Investigation for Baking. - SAS Detection and Investigation for Government. - SAS Detection and Investigation for Health Care. - SAS Detection and Investigation for Insurance. - SAS Detection and Investigation for Management.

3.4 Oracle

The Oracle analytics platform has the ability to offer all aspects of big data management to customers, especially scalability and real-time operation**.** Oracle provides business analytics solutions with machine learning (ML) and artificial intelligence (AI) in a single platform with cloud, self-service capabilities and advanced analytics. There are three main core components in Oracle analytics platform:

- Augmented analytics.
- Self-service analytics.
- Governed analytics. (Oracle 2019a.)

These three core components are discussed in the following section.

3.4.1 Augmented analytics – improving step

The Oracle Analytics Cloud uses Machine Learning to gather, explore and improve data in order to extract quick and accurate insights for the users. The solution can be used for natural-language generation (NLG) narratives and predictive analytics. (Oracle 2019a, 5.)

Data discovery

The Oracle Analytics Cloud applies intelligent Machine Learning to check the data first; after that, the application gives an explanation of the features or metric to the user to help them with the research. However, the results could be different for different users based on users' experiences or biases. For example, two users can access the same information, but they might get different explanations for the event. Smart augmentation for its data flows is an advanced technique of Oracle that other systems, such as Tableau, do not have. (Oracle 2019a, 5.)

Data enrichment

Intelligent Machine Learning can add extra data columns, which explain the metrics, and add more colours to the analytics interface. Furthermore, it also offers more methods of explaining, grouping or viewing the data to help users discover additional insights. (Oracle 2019a, 5.)

Mobile experience

Oracle applies natural language speech recognition with a smart recommendation engine to optimise the mobile experience for users. Thanks to Oracle Day by Day analytics app, the organisation can find out details about customer preferences, e.g. information about what, where, when the customers are interested in. This knowledge important to ensure that the organisation meet its customer expectations. Furthermore, it also helps the organisation identify potential customers or partners. (Oracle 2019a, 5)

Natural language generation-based narratives is the combination of visualisation and story context that help the users understand insights easier. (Oracle 2019a, 5)

Predictive analytics. A real-time operation is a standout feature of Oracle analytics cloud that allows users to enter data, click and get predictive analytics quickly (Oracle 2019a, 6).

3.4.2 Self-service analytics

Self-service data preparation authorises non-professional users to discover and utilise business analytics. In other words, the users can access, combine, improve and visualise data and then share it in a narrative. Moreover, the users are allowed to explore data visualisations and make their own analyses easily thanks to intuitive visual interaction. (Oracle 2019a, 6.)

Self-service data visualization

The users can create dashboards, diagrams, etc. by simple and powerful self-service tools. The tools are easy to use so the users do not need to know coding skills or IT help. (Oracle 2019a, 6.)

Smart SaaS integration

Oracle Analytics cloud combines smoothly with Oracle Cloud services and supports single sign-on and pass down security. Thus, that organization can avoid the time and setting up or maintaining security manually in many places. (Oracle 2019a, 6.)

Embedded insights

The users can gain insights, create comments and interactions and share them through to the organisation. This function facilitates a high level of interaction in the organisation. As a result, it allows the organisation to take advantage of insights in real-time. (Oracle 2019a, 6.)

Reliable, elastic cloud for security, performance and available

Oracle analytics cloud keeps developing and improving infrastructure and technology to ensure scalability, security, flexibility and reliable solution for customers (Oracle 2019a, 7).

3.4.3 Governed analytics

In order to generate insights from the data, users simply choose interactive visualisation and the application will establish a suitable or progressive computation. To address data security concerns and protect the organisation from data breaches, data visibility and access to content can be controlled by a role-based application. (Oracle 2019a, 7.)

Data virtualization

Data virtualization model takes a role in effective business insights. The application abstracts physical data from business view or knowledge by using prebuilt data virtualization. (Oracle 2019a, 7.)

Enterprise security

If the application provides engineers for the clouds, such as robust layered security, that could limit potential risks. The Oracle analytics provides Oracle Identify Cloud Service, an essential part of the Oracle Cloud structure. It has the ability to manage accession and core identity through a multitenant cloud platform. (Oracle 2019a, 7.)

Self-service dashboards

Oracle analytics cloud enables users to create dashboards to find the best model based on their specific purposes or business requirements. A good data visualisation, one that reveals important trends or insights, helps the company make an impact on strategy, risk management, decision making, etc. (Oracle 2019a, 7.)

Production reporting

The software is able to send a large amount of data, report quickly and directly to the users on demand or automatically. As a result, the organisation can skip the stages of inserting data and creating reports manually. This, consequently, saves time and lowers the workload of collecting and processing data. (Oracle 2019a, 7.)

4 PRACTICE

This section will discuss the implications of Big Data in real-world business operations with a special focus on the features Big Data analytics that can benefit Small and Medium-sized Enterprises (SMEs). SMEs as an essential part of the Finnish economy; studying the implications of Big Data analytics is thus highly relevant in the Finnish context.

4.1. SMEs in the European Union and Big Data

SMEs stands for Small and Medium-sized Enterprises. SMEs makes up around 99% of all business in the European Union (EU) and are a significant contributor to the EU's economy. In 2015, there were 23.4 million SMEs in the EU. Micro enterprises (i.e. enterprises with fewer than 10 employees) was 92.8% of all enterprises in the EU while large enterprises (i.e. enterprises with more than 250 employees) was only 0.2%. SMEs created jobs for 91 million people and contributed 3,934 billion euros in value. (European Commission 2018.) Hence, SMEs plays a very important role not only in the EU's economy but also in the world's economy. The opportunities and challenges of SMEs are described in the following parts.

4.1.1. Opportunities

To stand in the ever-competitive word's market nowadays, a company needs to catch and ride the trends, with big data being among those trends, to gain and retain its competitive advantages. The theory part of this thesis presents some ideas in which big data and big data analytics can benefit organizations (e.g. insights, customer behaviour analysis, predictive analytics, processing data, risk management, etc.). Big data provide a powerful tool to run businesses more efficiently. Such improvement in efficiency and productivity will result in lower operational costs, higher revenues, better profits, and more consistent and sustainable growth. In addition, SMEs can adapt to external changes better than larger organizations. Therefore, big data analytics can be implemented more easily and quickly in SMEs. Consequently, big data is very potential in helping SMEs create new businesses and achieve a higher level of productivity and innovativeness.

4.1.2. Challenges

According to a survey from Statista (2019), in 2017, a majority of SMEs were interested in digitalisation. Around 25% of SMES mentioned that digitalisation has a strong role in the organisations. Moreover, digitalisation is an essential feature of the product or service offerings for 22% of SMEs in the survey. Nevertheless, only a small proportion of SMEs (more than 30%) claimed that they used big data or analytics. (Statista 2019.) The following part will describe a few barriers that prevent SMEs from the implementation of big data analytics. Those barriers are discussed in the following sections.

Lack of understanding. The e-skill survey in the UK and the survey among German SMEs demonstrate that only 30% to 40% SMEs understand big data analytics. Many of SMEs still doubt the benefits for companies from big data and their data do not qualify for at least one of the big data dimensions (3V's). (Coleman, Göb, Manco, Pievatolo, Tort-Martorell & Reis 2016.)

Dominance of domain specialist. Companies tend to focus on developing their own field instead of expanding to other fields. Hence, they do not pay much attention to the new trend 'Big Data' if they do not have enough experiences in the big data field. (Coleman et al. 2016.)

Cultural barriers and intrinsic conservatism. SMEs either do not have a big interest in big data or have poor infrastructure. If a company does not pay attention to properly setting up the system from the beginning, it will end up with poor data infrastructure. This poor infrastructure will, in turn, become a hinderance when the company wants to invest in big data. With poor infrastructure, data will likely be in an incorrect format and consequently cannot be accessed and processed easily. (Coleman et al. 2016.)

Shortage of in-house data analytic expertise. Firstly, at the moment, the number of highly skilled data analysts are not enough to meet the demand, which is getting higher because of the growing prevalence and popularity of big data analytics. Secondly, because of the high cost of setting up big data infrastructure, many SMEs are not sure if the added benefits generated by big data analytics is worth the investment. Thus, investments in big data infrastructure is perceived as risky, especially by SMEs with few employees. (Coleman et al. 2016.)

Shortage of useful and affordable consulting. Normally, big data analytics programs are developed by large corporations for big organizations while most SMEs seek accounting, hardware/ software or

IT-related services at a lower price range. Due to the lower price range, big data analytics for the SMEs sector does not get much attention from developers. In addition, the price for business analytics programs is still too high for many SMEs. (Coleman et al. 2016.)

Non-transparent software market. For SMEs with little or no expertise, choosing suitable analytics software is difficult as there is still a lack of transparency that allows the comparison and evaluation across different analytics software programs. (Coleman et al. 2016).

Lack of intuitive software. The technical complexity of business analytics is generally considered as a spectrum with the following two extremes: potential analytics vs simple analytics. (Coleman et al. 2016). The characteristics of these two extremes are illustrated in Table 5. It is worth noting that potential analytics lacks intuitiveness and requires IT and data analytics expertise to implement. This poses a potential barrier for SMEs without IT/data analytics expertise from big data analytics implementation.

TABLE 5. Potential analytics vs Simple analytics (developed from Coleman et al. 2016)

Potential analytics	Simple analytics
• High quality performance • Complicated to use • Required IT expert, data analyst	• Less quality performance • Easy to use • Do not require data analyst, IT expert

Concerns on data security. Due to poor infrastructure, the lack of IT experts, and a lower budget compared to large organizations, SMEs tend to be easier targets of cyber-attacks. Therefore, this could be the biggest barrier between SMEs and big data analytics implementation. In some cases, the costs and risks of data analytics implementation outweigh its the potential benefits. (Coleman et al. 2016.)

Concerns about data protection and data privacy. Customer data processing and analyses must abide the law of data protection and privacy. According to European Commission (2019), since May 2018, General Data Protection Regulation was mandated to all companies in the EU (European Commission 2019). There are not many SMEs that can afford to hire a lawyer with enough expertise to support the organisation to understand all the rules and requirements of the regulation. (Coleman et al. 2016.)

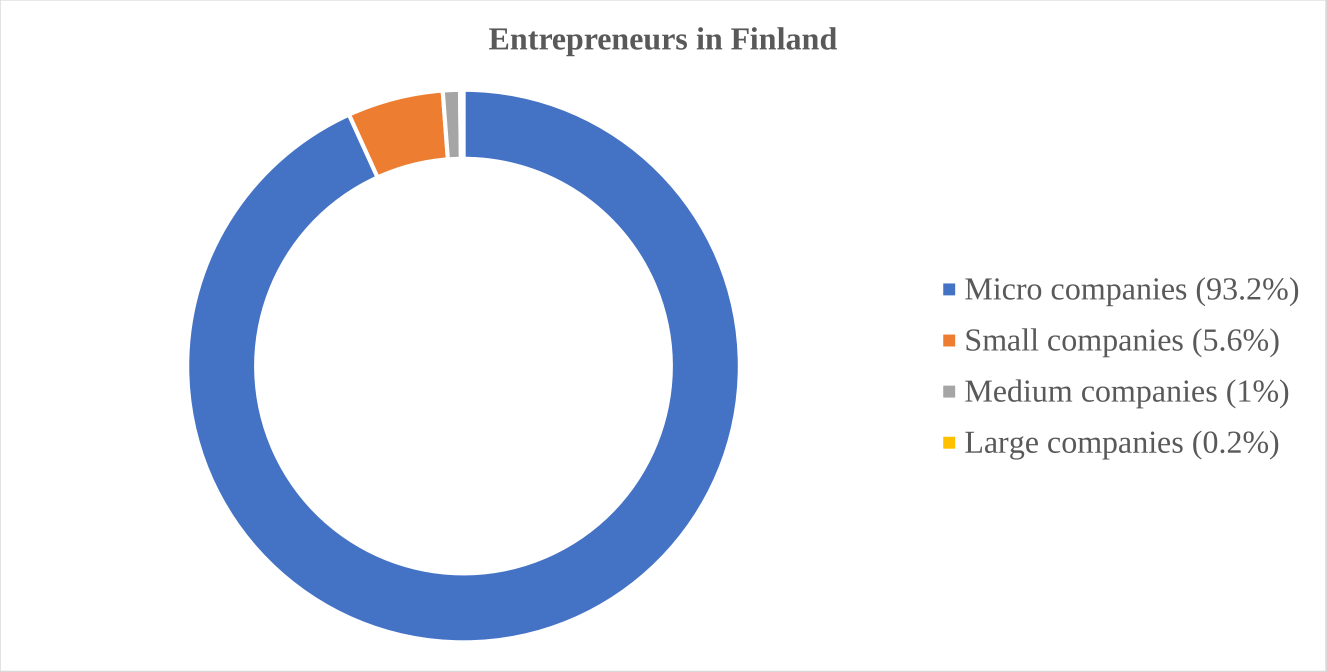
4.2. SMEs in Finland

According to Suomen Yrittäjät (2019), in 2017, there are a total of 286,934 companies in Finland excluding agriculture, fisheries and forestry. Micro companies constitute the biggest proportion, 93.2% (267,447 companies), while large companies account for the smallest proportion of only 0.2% (615 companies). The second biggest category is small companies at 5.6% and the third biggest category is medium companies at 1% (See Table 6 and Graph 1).

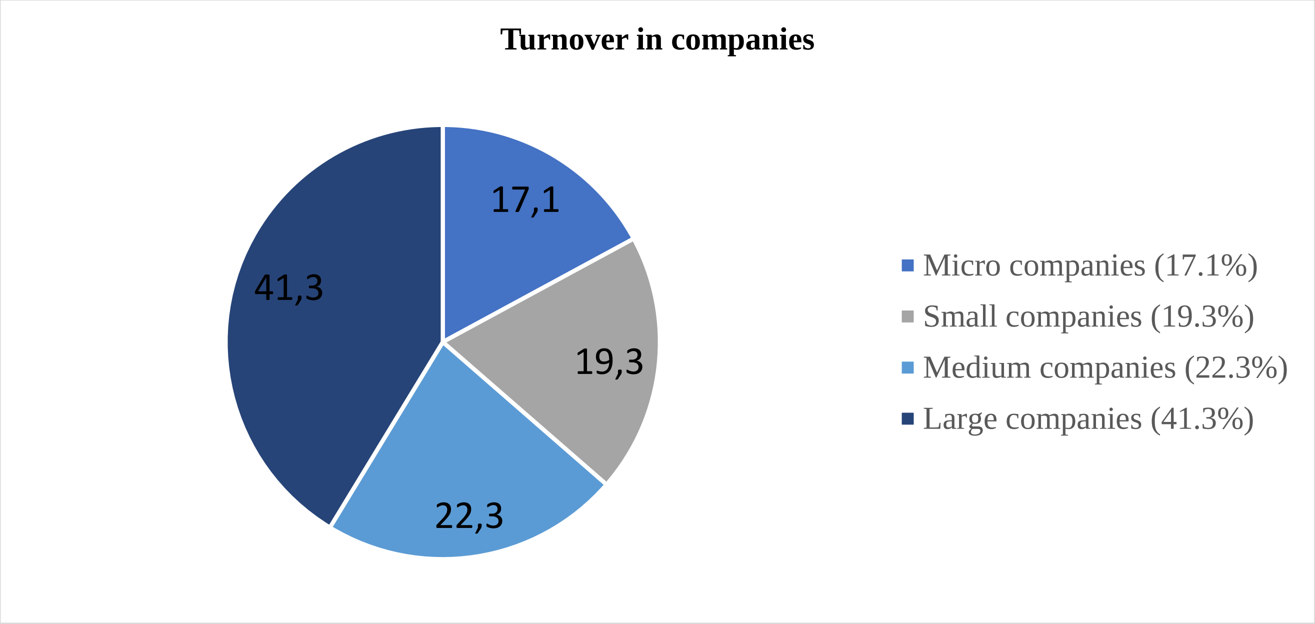
TABLE 6. Classification of Finnish companies based on size

Category	Count	Percentage
Micro companies (1-9 employees)	267447	93.2%
Small companies (10-49 employees)	15989	5.6%
Medium companies (50-249 employees)	2883	1%
Large companies (>250 employees)	615	0.2%
Total	286934	

The turnover of all enterprises was 409 billion euro in 2017. Large companies generate 169 billion euros, constituting 41% of the total revenue. Among the other 59%, medium companies accounted for 22% micro companies accounted for 17%, and small companies accounted for a mere 2%. Based on the above information, SMEs contributed almost 60% turnover while large companies contributed only round 40%. Thus, SMEs play a significant role in the Finnish economy (See Graph 2).



GRAPH 1. Entrepreneurs in Finland (developed from Suomen Yrittäjät)



GRAPH 2. Turnover in companies (developed from Suomen Yrittäjät)

Centria University Applied Sciences has many partners around the Central Ostrobothnia region that are SMEs, e.g. Accuser Oy, Ojala group, Scanfil Oy, or Sievi tools Oy. For this reason, if Centria integrates more teaching and research on the topic of big data analytics, it could generate potential benefits for the partners.

4.3. Comparing features of different software programs

This part discusses aspects of data security and Internet of Things that companies, especially SMEs, should consider before purchasing and installing data analytics software programs. It is important for organisations and employers to choose appropriate software programs that are suitable for the intended purposes and the company's concepts and operations philosophy. Additionally, when searching for big data analytics software programs to purchase, companies should also take into consideration the user-friendliness of the software programs; this is especially important for SMEs that lack IT expertise and/or budget for investment in big data.

4.3.1. Data Security

Even though cloud computing is one of the enablers for numerous innovative disruptions, such as IoT and Industry 4.0, the growing prevalence of cloud computing raises security concerns among its users. Users of cloud computing are generally subjected to a higher threat of cyber-attacks compared to the traditional way where data are stored locally. Cyber-attacks on important data storages can have serious consequences. No organisation wants important or sensitive information to be leaked from the systems; to the extreme, data leaks could completely destroy an organisation. Furthermore, leaks of external information, such as customer information, can also have some complicated legal and ethical implications. The following parts describe how the previously-described analytics software programs address the data security concerns.

Pentaho

Pentaho provides Hitachi Data Instance Director (HDID). The functionalities of HDID (See Figure 10) are described as follows.

- Making a copy of production data without interruptions. In case some error happens, there is still a copy version data to ensure no data loss.
- HDID controls user access by role-based access. This means that HDID can protect unexpected user access to the information are not suitable or belong to the category of limited-access data. HDID can support and clarify unique profiles.
- Backup data cannot be deleted or changed by modifying.

- "Keep pace with the demands of critical application" is a solution HDID can offer while traditional methods cannot. (Hitachi 2019a.)

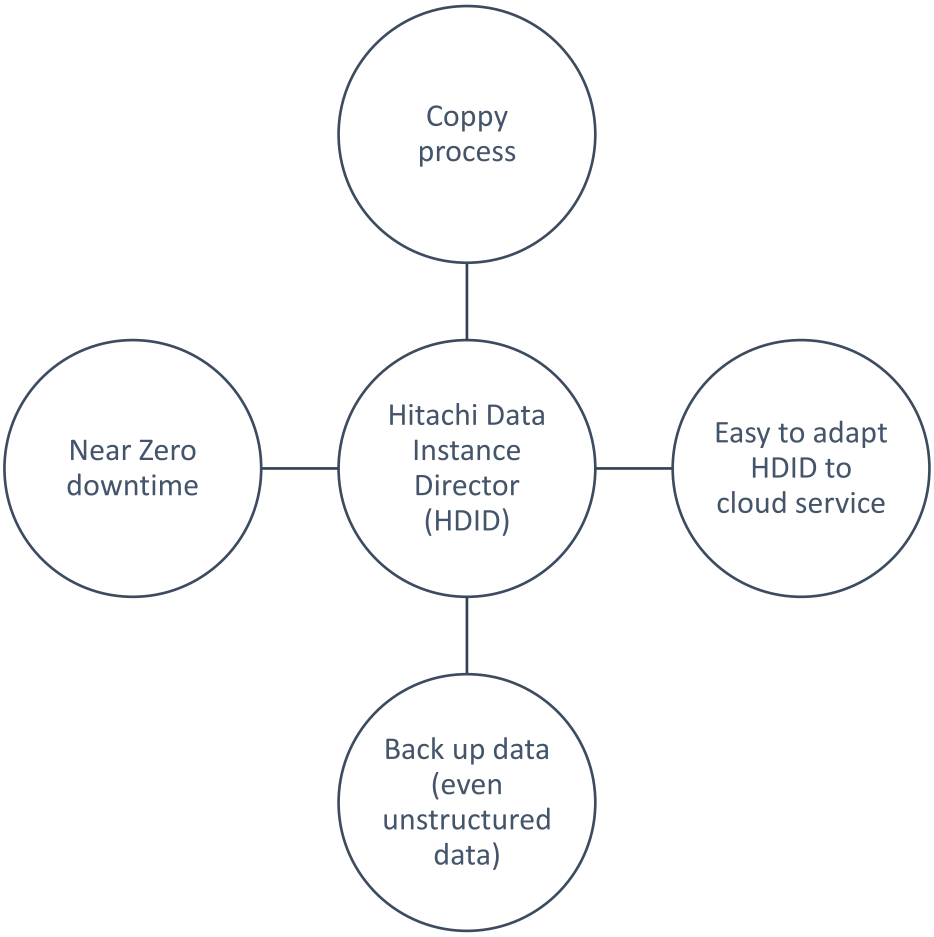

Coppy process
Near Zero downtime
Hitachi Data Instance Director (HDID)
Easy to adapt HDID to cloud service
Back up data (even unstructured data)

FIGURE 10. Hitachi Instance Director (HDID) (developed from Hitachi 2019d)

Oracle

Oracle Data Security comes with four priorities. It can prevent data lost, detect suspicious activities, evaluate security postures and apply data access control at the source by data driven security (See Figure 11, 12 and 13). (Oracle 2019c.)

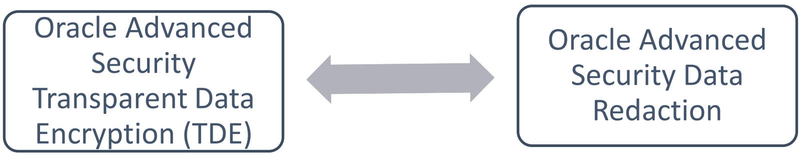

Oracle Advanced Security Transparent Data Encryption (TDE)
Oracle Advanced Security Data Redaction

FIGURE 11. Oracle Advanced Security (developed from Oracle 2019c)

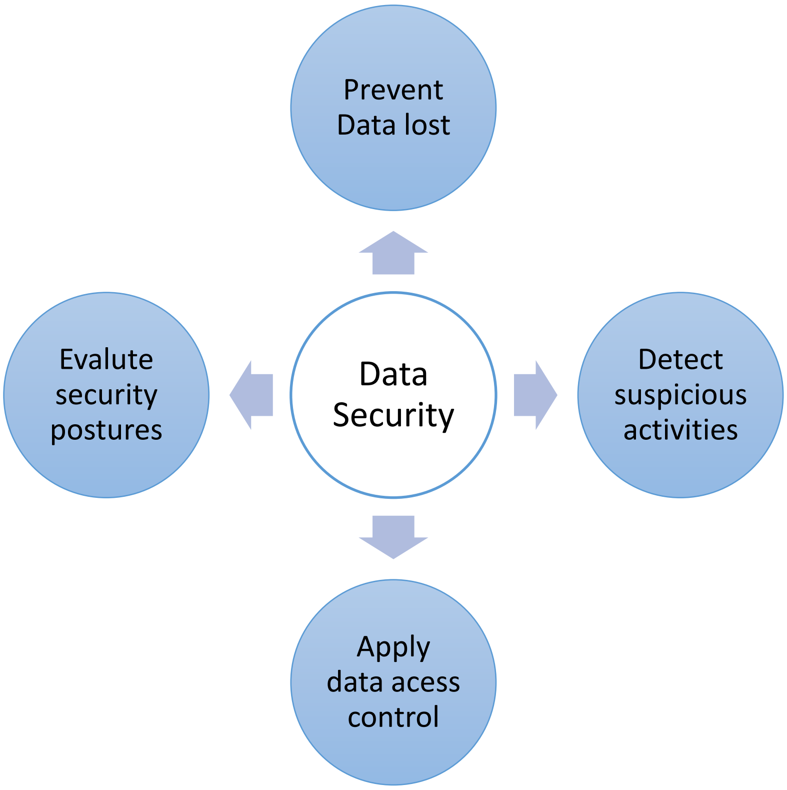


FIGURE 12. Oracle Data Security (developed from Oracle 2019c)

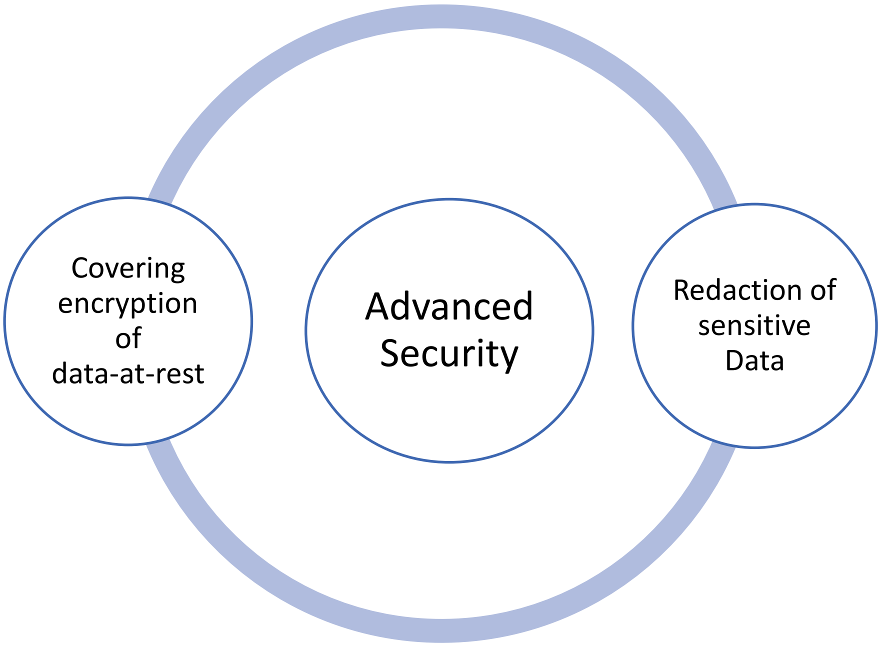

Covering encryption of data-at-rest
Advanced Security
Redaction of sensitive Data

FIGURE 13. Oracle Advanced Security Operation (Oracle 2019C)

The purpose of Oracle Advanced Security Transparent Data Encryption (TDE) is to prevent attacks trying to bypass the database and reading important information from backup or exported database or data files at operation level. TDE encrypts the database layer and therefore only authenticated users are allowed to access the database and important information. As a result, unauthenticated users will not be able to access the databases or read sensitive data in clear text. (Oracle 2019d, 6). Furthermore, Oracle Advanced Security Data Redaction organises important data in query results before data leave the database and can be seen by authenticated users. (Oracle, Encryption redaction Oracle advance security, 29th August 2019, p. 11)

SAS

SAS Cybersecurity Solution provides the following benefits (See Figure14).

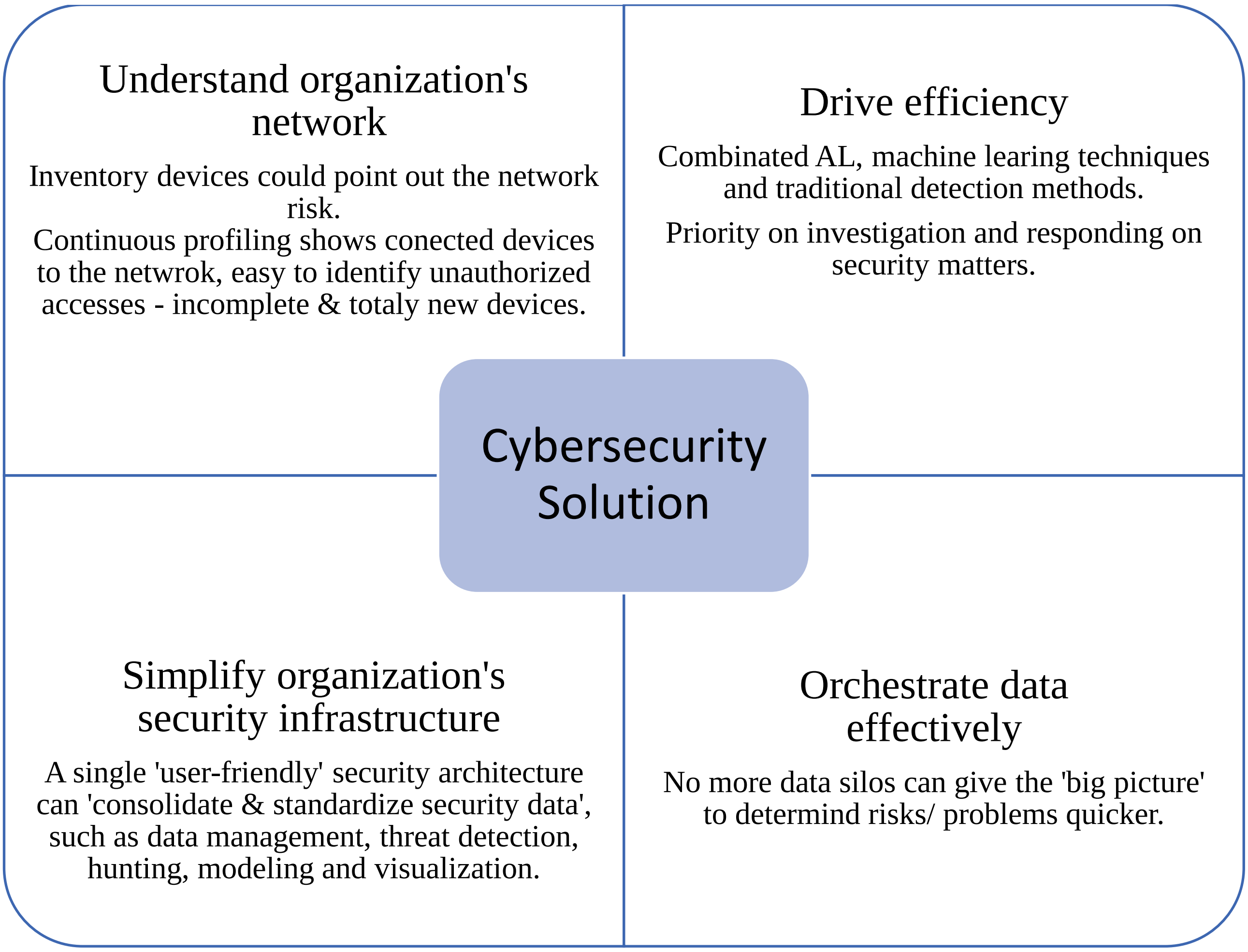
- More than seventy device analytics determine and analyse network activities.

- Out of the box, a combination of NetFlow, web proxy, DHCP, DNS, authentic and endpoint data to create external analytics.

- Those two features are able to determine/detect internal and external risks.

- Reduce the time to detect and respond to unusual behaviour/ activities.

- SAS Cybersecurity Solution gets benefits from SAS Viya. (SAS 2019a.)



Understand organization's network

Inventory devices could point out the network risk.
Continuous profiling shows conected devices to the netwrok, easy to identify unauthorized accesses - incomplete & totaly new devices.

Combinated AL, machine learing techniques and traditional detection methods.
Priority on investigation and responding on security matters.

Simplify organization's security infrastructure

A single 'user-friendly' security architecture can 'consolidate & standardize security data', such as data management, threat detection, hunting, modeling and visualization.

No more data silos can give the 'big picture' to determind risks/ problems quicker.

FIGURE 14. SAS Cybersecurity Solution (developed from SAS 2019a)

4.3.2. Internet of things (IoT)

The concept of ’Internet of things’ refers to everyday objects, such as household appliances, vehicles, machines, etc. that contains electronics and built-in devices for Internet access. This Internet connection enables them to be controlled, to transmit, and to share data online. There are many possible areas where IoT solutions can be applied both in the private and commercial spheres. (Oracle IoT for connected worksites & intelligent maintenance. 2016). The following sections discuss how the analytics software programs address IoT in their functionalities.

Pentaho

Pentaho IoT Lumada platform also focuses on reducing risk, shortening time to value, providing quick outcome, improving performance by process and analysing data in real-time (See Table 7).

TABLE 7. Ability of Pentaho's Industrial Internet of Things

<table>
<tr><th colspan="4">Ability of Industrial Internet of Things (IIoT)</th></tr>
<tr><td>Scheduled production line based on analysis.</td><td>Managed production line and maintained the engine at the same time.</td><td>Controled quality and quality of products by adapting suitable route and velocity in production line.7x</td><td>Combination of real-time analysis, visualizaztion to minimum inccident.</td></tr>
<tr><td colspan="4"></td></tr>
</table>

Lumada IoT platform architecture is divided into six significant layers: Foundry, Edge, Core, Data Management, Analytics and Studio (See Figure 15 and 16).

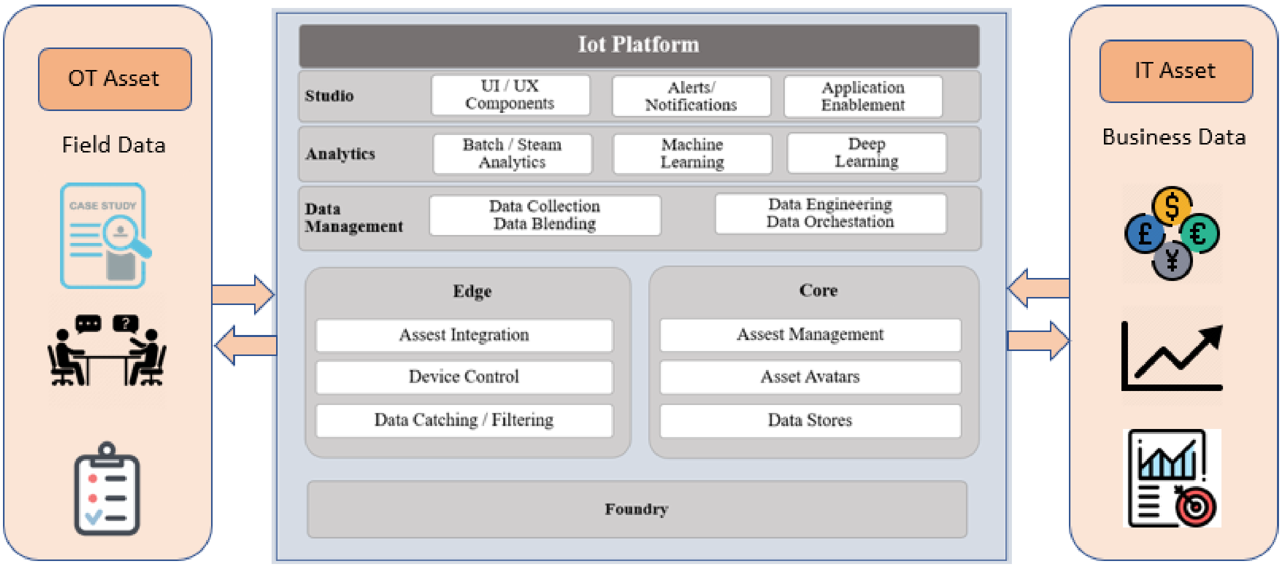
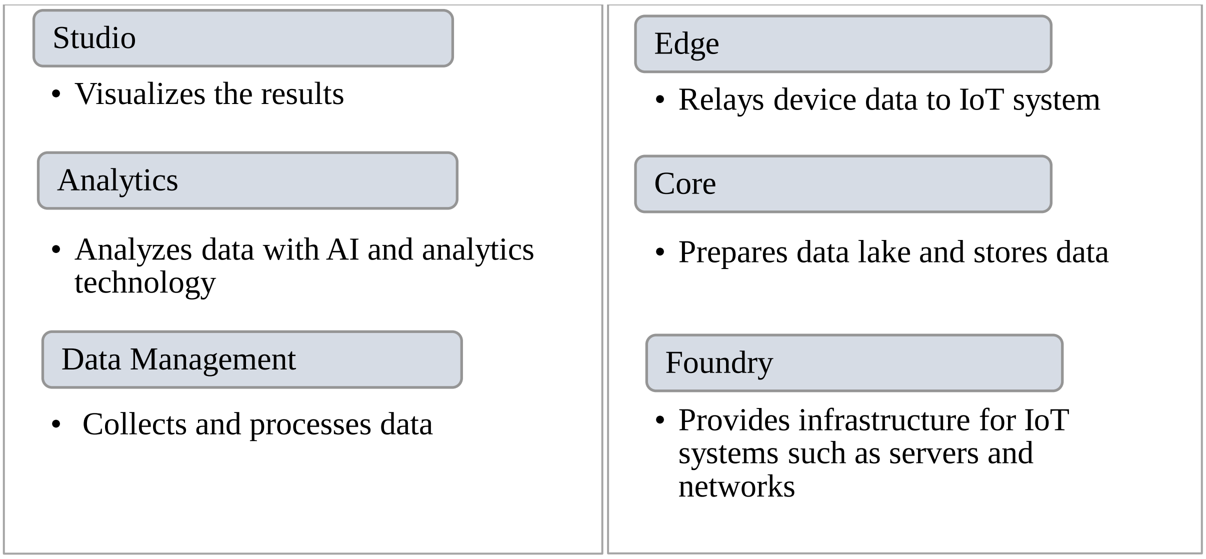


FIGURE 15. Lumada IoT platform architecture (developed from Hitachi 2019b)



Studio

- Visualizes the results

Analytics

- Analyzes data with AI and analytics technology

Data Management

- Collects and processes data

Edge

- Relays device data to IoT system

Core

- Prepares data lake and stores data

Foundry

- Provides infrastructure for IoT systems such as servers and networks

FIGURE 16. Six architecture components of Lumada IoT (developed from Hitachi 2019b)

Oracle

Oracle IOT can support any organisation in planning, building, and operations. For instance, the occupational safety aspects of *Construction and Engineering* operations can be tracked and controlled with sensors. More specifically, information about temperature, air quality, humidity, noise, proximity, etc. can be measured by sensors. Subsequently, the information can be sent to the work site on the Oracle system. The Oracle system will detect and send notifications/warnings if any measures exceed the acceptable range of their respective quality standards. With timely warnings, managers can solve the problems quickly.

IoT support management

The Oracle worksite also enables a controlled operation. The streamlined information and real-time data keep managers updated about what is happening with one or more projects. For example, the work site can display information about e.g. materials, employees, etc. Such information could help managers with decisions regarding scheduling or material management.

IoT can also track information about the performance measures of machineries; it can provide reports on performance measures in real-time or warnings when an unusual occurrence takes place. Consequently, this allows companies to know when it is necessary to maintain the engines. (Oracle 2016.). In addition, Oracle IoT can be applicable and useful in many different sectors, such as retail, healthcare, industrial manufacturing, construction and engineering and etc.

SAS

SAS offers businesses and organisations with analytics solutions for IoT-enabled data ranging from machine learning, deep learning to artificial intelligence. The solutions are to help organisations transform the vast amount of data into valuable information. The information can then be used in statistical, descriptive, and predictive analyses. SAS promotes its use of AI in IoT analytics solutions and calls it AIOT (Artificial intelligence of things). The AI aspect of IoT analytics is described to improve interactions and relations between customers and companies by e.g. making it possible to predict maintenance, optimise allocation of resources, and offer customised promotions and discounts to customers. (SAS 2019c.)

5 CONCLUSION

Big data analytics is a significant and extensive field. The demand for big data analytics is increasing thanks to the various benefits it has to offer to organisations in various industry segments. Besides the evident opportunities associated with big data analytics, there are numerous challenges that organisations have to overcome to implement and execute big data analytics. As an answer for such challenges, there are various analytics software programs for organisations in the market. This thesis discusses three of the most popular data analytics software programs, namely SAS, Oracle, and Pentaho. In general, a common goal across all software companies is to provide suitable products for customers.

There are a few considerations when it comes to choosing a suitable system. First of all, companies should choose products and solutions with high security functions to protect data about customers and other members of the organisations. That is the most important consideration for companies when choosing a data analytics system. Secondly, companies should also consider the user-friendliness aspect of the software programs. This is especially important for SMEs, as they often lack the technical expertise. User friendly software programs would allow SMEs to execute big data analytics without an IT or data analytics expert.

Thirdly, companies should also consider the costs of the analytics software program against the potential benefits that can be generated from implementing the program. The costs include purchasing costs, implementation costs, costs of updates, etc. A careful cost-benefit analysis would allow companies to make better purchasing decisions. It also allows companies to choose a suitable software for their needs. Lastly, companies should consider the Internet of Things aspect of the analytics software program. The IoT aspect of SAS, Oracle and Pentaho are quite similar in nature. All three programs focus on real-time analytics, e.g. creating schedule for production process, managing the production line, etc., in order to create a smooth production line and reduce incidents. In the long run, those functions enable companies to reduce operational costs, reduce waste, improve productivity, and improve efficiency.

After a thorough analysis of the three analytics software application, this thesis would recommend Oracle for SMEs. Besides its functionalities, Oracle also provides free online training and certification. As the result, everyone can access to courses and get certifications. This is extremely beneficial for SMEs, especially for those that lack in-house expertise; essentially, Oracle offers free training to organisation in addition to the technical functionalities. In conclusion, choosing a suitable analytics software program for SMEs is a significant yet challenging process; organisations should carefully study different applications to identify the best fit for their specific needs and expectations.

REFERENCES

Ahmed, O., Fatima-Zahra, B., Ayoub, A, L. & Samir,B. 2017. Big Data technologies: A survey. Available: https://www.sciencedirect.com/science/article/pii/S1319157817300034. Accessed 3 August 2019.

Amir, G. & Murtaza, H. 2014. Beyond the hype: Big data concepts, methods, and analytics. Available: https://www.academia.edu/39955757/Beyond_the_hype_Big_data_concepts_methods_and_analytics-NC-ND_license_http_creativecommons.org_licenses_by-nc-nd_3.0. Accessed 17 July 2019.

Anasse, B., Mohamed, C. & Tommy, J. 2017. Predictive Analytics for dummies.

Andrew, M. & Erik, B. 2012. Harvard Business Review. Big Data: The Management Revolution. Available: https://www.academia.edu/2662701/Big_Data_The_Management_Revolution. Accessed 24 September 2019.

Chun-Wei, T., Chin-Feng L., Han-Chieh C., & Athanasios V.V. 2015. Big data analytics: a survey. Available: https://journalofbigdata.springeropen.com/track/pdf/10.1186/s40537-015-0030-3. Accessed 2 July 2019.

Danyel, F., Rob, D., Mary, C., & Steven, D. 2012. Interactions with Big Data Analytics. Available: https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/inteactions_big_data.pdf. Accessed 14 October 2019.

David, R., John, G. & John, R. 2018. The Digitization of the World From Edge to Core. Available: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. Accessed 16 October 2019.

Doug, B., Sanjeev, K., Harry, Li., Jason, S. & Peter, V. 2012. Finding a needle in Haystack: Facebook’s photo storage. Available: https://www.usenix.org/legacy/event/osdi10/tech/full_papers/Beaver.pdf. Accessed 12 June 2019.

European Commission. 2019. EU data protection rules. Available: https://ec.europa.eu/info/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules/eu-data-protection-rules_en. Accessed 18 October 2019.

European Commission. 2018. Small and medium-sized enterprises: an overview. Available: https://ec.europa.eu/eurostat/web/products-eurostat-news/-/EDN-20181119-1. Accessed 15 September 2019.

Hisham, A., John, S., Franz, K., Emily, S., Kym, P. 2011. Semantically Enhanced Information Extraction.

Hitachi. 2018. Pentaho Business Analytics Platform Ingest, Prepare, Blend and Analyze. Available: https://www.hitachivantara.com/en-us/pdfd/datasheet/pentaho-business-analytics-platform-datasheet.pdf. Accessed 20 October 2019.

Hitachi. 2019a. Hitachi Data Instance Director – Datasheet. Available: https://www.hitachivantara.com/en-us/pdfd/datasheet/data-instance-director-datasheet.pdf. Accessed 28 October 2019.

Hitachi. 2019b. Lumada Catalog. Available: https://www.hitachi.com/products/it/lumada/global/en/download/data/lumada_catalog.pdf

Accessed 2 April 2020.

Hitachi. 2019c. Pentaho and Machine Learning Orchestration. Available: https://www.hitachivantara.com/en-us/pdfd/datasheet/pentaho-machine-learning-orchestration-datasheet.pdf. Accessed 25 October 2019.

Hitachi. 2019d. Pentaho Data Integration The Power to Access, Prepare and Blend Multiple Data Sources Faster. Available: https://www.hitachivantara.com/en-us/pdfd/datasheet/pentaho-data-integration-datasheet.pdf. Accessed 18 October 2019.

Judith, H., Alan, N., Dr. Fern, H., & Marcia, K. 2013. Big data for dummies.

Kan, L., Lin, Z., Heyan, H. 2018. Social Influence Analysis: Models, Methods, and Evaluation. Available: https://www.sciencedirect.com/science/article/pii/S2095809917308056. Accessed 15 July 2019.

Kai, Z., Tomasa, M., Talal, R., Marcin, W., Yevgeniy, V. 2018. Attacking Similarity-Based Link Prediction in Social Networks. Available: https://arxiv.org/pdf/1809.08368.pdf. Accessed 21 August 2019.

Kiron, K. 2012. Video Analytics – Choice of Architecture. Available: http://www.norikkonsult.com/pdf/VideoAnalyticsChoiceofArchitecture.pdf. Accessed 12 November 2019.

Luis, V., Sergio, S., Reinier, G., Anibal, A., Miguel, P. & L. Enrique, S. 2017. A Cloud-based architecture for smart video surveillance. Available: https://www.semanticscholar.org/paper/A-CLOUD-BASED-ARCHITECTURE-FOR-SMART-VIDEO-Valentin-Serrano/afad2ca47366c194007fc50a79ed3735cb2119ff. Accessed 19 July 2019.

Michele, C., Fosca, G., Dino, P. 2012. A Classification for Community Discovery Methods in Complex Networks. Available: https://arxiv.org/abs/1206.3552. Accessed 15 September 2019.

Mohammed J. Zaki. 2011. Link prediction in social networks. Available: https://www.researchgate.net/publication/226566834_A_Survey_of_Link_Prediction_in_Social_Networks. Accessed 28 September 2019.

Mohsen, J., Hassan, A. 2016. Different Aspects of Social Network Analysis. Available: https://www.cs.sfu.ca/~oschulte/teaching/socialnetwork/papers/SNA-intro-mohsen.pdf. Accessed 28 July 2019.

Nada, E., Ahmed,E. 2016. Big Data Analytics in Support of the Decision-Making Process. Available: https://www.sciencedirect.com/science/article/pii/S1877050916324206. Accessed 13 September 2019.

Oracle. 2013. Oracle: Big Data for the Enterprise. Available: http://www.oracle.com/us/products/database/big-data-for-enterprise-519135.pdf. Accessed 8 July 2019.

Oracle. 2018. Encryption Encryption and Redaction in Oracle Database 18c with Oracle Advanced Security. Available https://www.oracle.com/a/ocom/docs/solutions/business-analytics/oracle-analytics-compete-to-win-white-paper.pdf. Accessed on 18th November 2019.

Oracle. 2019a. A Technical Buyer's Guide to Analytics. Available: https://explore.oracle.com/c/oracle-analytics-com?x=nch80j&xs=82580. Accessed 23 July 2019.

Oracle. 2019b. Oracle Analytics Cloud Augmented Analytics: Start the Journey! Available: http://www.oracle.com/us/solutions/oracle-analytics-cloud-3711629.pdf. Accessed 23 July 2019.

Oracle. 2019c. Oracle Advance Security. Access 2nd July 2019. Available: https://www.oracle.com/a/tech/docs/dbsec/aso/advanced-security-ds-19c.pdf. Accessed 15 October 2019.

Oracle. 2019d. Oracle Encryption Redaction Oracle Advanced Security. Accessed 9 September 2019.

Oracle IoT for connected worksites & intelligent maintenance. 2016. Video on YouTube. Available: https://www.youtube.com/watch?v=wKuyc_tU9vI. Accessed 25 November 2019.

Paul, C.Z., Chris, E., Dirk, D., Thomas, D., George, L. 2012. Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data.

Philip, R. 2013. Managing Big Data. Available: http://www.datascienceassn.org/sites/default/files/Managing%20Big%20Data%202013.pdf. Accessed 28 June 2019.

SAS. 2012. Big Data Meets Big Data Analytics. Available: https://www.academia.edu/35437586/Big_Data_Meets_Big_Data_Analytics_Three_Key_Technologies_for_Extracting_Real-Time_Business_Value_from_the_Big_Data_That_Threatens_to_Overwhelm_Traditional_Computing_Architectures. Accessed 9 October 2019.

SAS. 2019a. SAS Cybersecurity._ Available: https://www.sas.com/content/dam/SAS/documents/product-collateral/fact-sheets/en/sas-cybersecurity-110294.pdf. Accessed 15 September 2019.

SAS. 2019b. SAS Viya Built for innovation so you can meet your biggest analytical challenges. Available: https://www.sas.com/content/dam/SAS/en_us/doc/overviewbrochure/sas-viya-108233.pdf. Accessed 7 November 2019.

SAS. 2019c. IOT SOLUTIONS Advancing toward the artificial intelligence of things - AIoT. Available: https://www.sas.com/en_au/solutions/iot.html. Accessed 25 October 2019. Accessed 23 September 2019.

Shirley, C., Rainen, G., Giueppe, M., Antonio, P., Xavier, T., Macro,S, R. 2008. How Can SMEs Benefit from Big Data? Challenges and a Path Forward._Available: https://www.academia.edu/35085839/How_Can_SMEs_Benefit_from_Big_Data_Challenges_and_a_Path_Forward. Accessed 3 August 2019.

Sonit, S. 2018. Natural Language Process for Information Extraction. Available: https://www.researchgate.net/publication/326264437_Natural_Language_Processing_for_Information_Extraction. Accessed 29 September 2019.

Statista. 2019. State of digitalization in SMEs in Finland 2017. Available: https://www.statista.com/statistics/881848/digitalization-state-in-smes-in-finland/. Accessed on 1 October 2019.

Suomen Yrittäjät. 2019. Finland cannot be renewed without enterprise. Available: https://www.yrittajat.fi/sites/default/files/sy_esittely2019_en-gb.pdf. Accessed 14 July 2019.

Tableau. Visual analysis best practice: A guidebook. Available: https://www.tableau.com/learn/whitepapers/tableau-visual-guidebook. Accessed 8 July 2019.

The Economist. 2010. Data, data everywhere. Available: https://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf. Accessed 28 July 2019.

Tom, W. 2012. Hadoop: The Definitive Guide, 3rd edition.

Vernon, T. 2016. IDC white paper Reducing the Time to Value for Internet of Thing Deployments. Available: https://www.oracle.com/webfolder/s/delivery_production/docs/FY16h1/doc25/IDCWhitePaperFinal.pdf. Accessed 21 September 2019.

Wei-Dong, Z., Bob, F., Daniel, G., Vijay, G., Josemina, M., Amarjeet, M., Tetsuya, N., Mark, P., Jane, S. & Martin, T. 2014. IBM Watson Content Analytics: Discovering Actionable Insight from Your Content. Available: https://books.google.se/books/about/IBM_Watson_Content_Analytics_Discovering.html?id=L1j3AwAAQBAJ&printsec=frontcover&source=kp_read_button&redir_esc=y#v=onepage&q=natural&f=false. Accessed 27 October 2019.