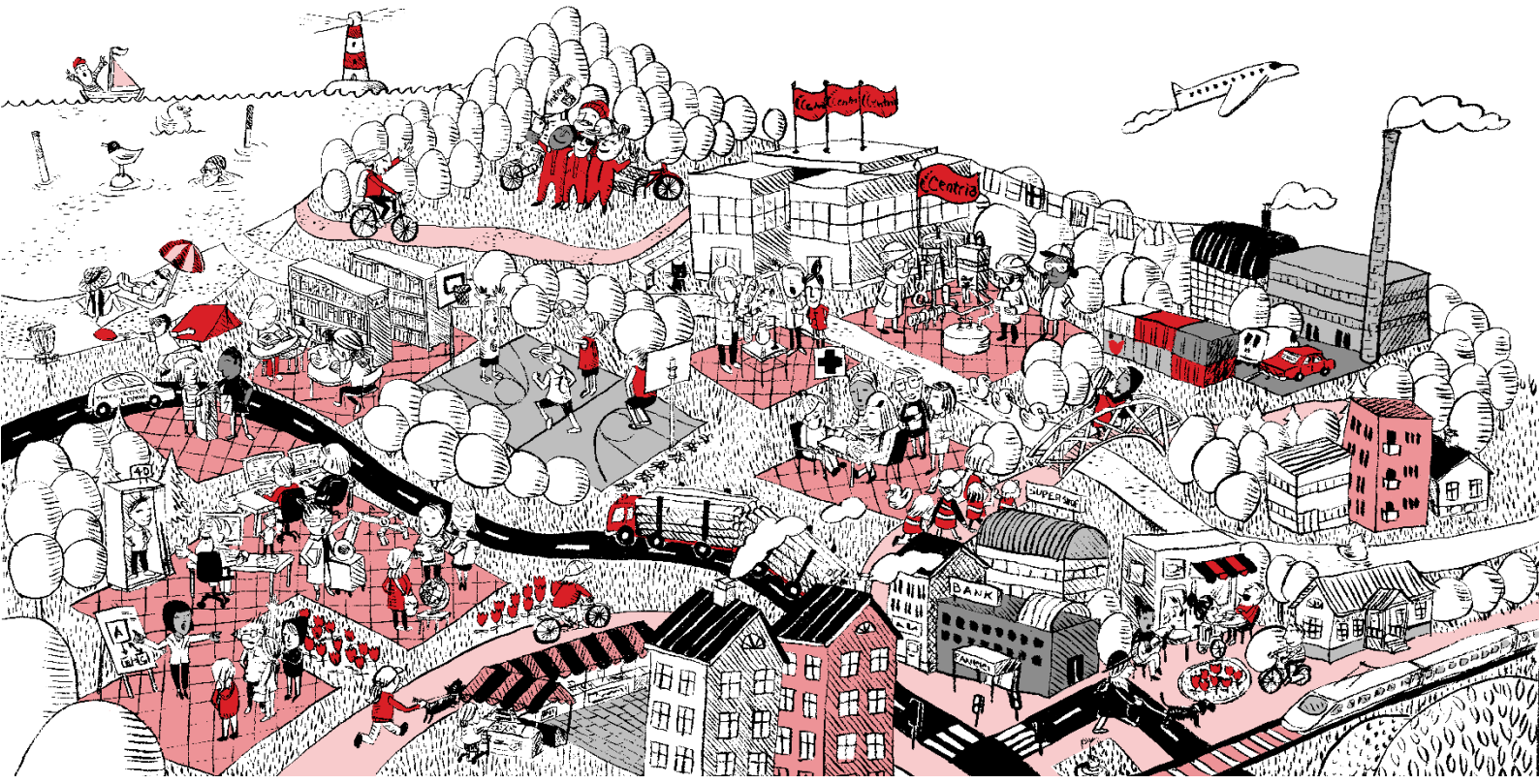
Don Duma

RECOGNIZING THE VALUE OF DATA IN BUSINESS OPERATIONS

A study on gathering internal and external data and ways to utilize it in business strategies

Thesis
CENTRIA UNIVERSITY OF APPLIED SCIENCES
Business Management, Enterprise Resource Planning (ERP)
October 2021



ABSTRACT

Centria University of Applied Sciences	Date October 2021	Author Don Duma
Degree programme Business Management, Enterprise Resource Planning (ERP)		
Name of thesis RECOGNIZING THE VALUE OF DATA IN BUSINESS OPERATIONS. A study on gathering internal and external data and ways to utilize it in business strategies		
Centria supervisor Janne Peltoniemi		**Pages** 47 + 7

The purpose of the study was to discover how organizations in the regions of Pietarsaari and Kokkola in Finland collect data and how data are utilised to make accurate and informed decisions to gain and maintain competitive advantage. The study was based on a questionnaire survey which was sent to sixty-three organizations' leaders operating in different business sectors. Moreover, interviews were conducted with some of the leaders to determine how their organizations gather data, how they process it, and how the insights gained influence business operations and decision-making processes.

The qualitative method was adopted for the research and the questionnaire contained fifteen strict and unstructured questions which were sent to leaders and their results were compared, analysed and findings were made. The survey focused on collection and usage of internal data, collection and usage of external data, and the concepts used to store the collected data before it is processed and analysed. Therefore, data storage solutions were considered and thus, the concept of data repository such as data lake technology was examined. However, different data structures models were out of scope of the study.

The findings of the research clearly indicated that most organizations that responded collect various data from various sources. The sources were both internal and external, and different applications, sensors, and devices were the collectors of data. However, the storage of data and its utilisation varied from organization to organization. Furthermore, the data collected was utilised differently depending on the business sector and size of the organization.

Finally, the majority of respondents admitted that their organizations are dependent on insights from collected data and will be committed to the gathering of data in the future.

Key words
Cloud storage, data collection, data cubes, data lake, data marts, data repository, data storage, data warehouse, external data, internal data, metadata, structured data, unstructured data

CONCEPT DEFINITIONS

CRM

(Customer relationship management) is a software that helps attract new customers and develop customer loyalty and is important in the retention of existing customers.

ERP

(Enterprise Resource Planning) is an information system used to plan and manage all the core processes of an organization.

IOT

(Internet Of Things) refers to network of physical objects (devices) with purpose of communicating and exchanging data over the internet.

OLTP

(Online transactional Processing) according to Microsoft refers to the management of transactional data using computer (Microsoft corporation, 2021).

UCI

The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms.

ABSTRACT
CONCEPT DEFINITIONS
CONTENTS

FIGURES

TABLES

1 INTRODUCTION

The topic of the study is recognizing the value of data in business operations, a study on gathering internal and external data and ways to utilize it in business strategies. The study will endeavour to deal with the complex situation of various kind of data currently collected within organizations. The aim of the study is to discover if organizations are aware that they are collecting different kinds of data. Additionally, the sources which are used to collect data will be discussed. The data collected will be divided in two categories depending on the source provenance and finally the repository used to store data will be considered.

The background of the study derives from the fact that the usability of information technology (IT) and information systems have become an integral part of any modern organization irrespective of the business sector. Organizations have different smart devices and applications installed on their premises, computer aided devices and sensors are incorporated on operating machines, websites have cookies for data collections and monitoring, smartphones and other handheld devices register activities, ERP (enterprise resource planning) systems, and many more that constantly collect data.

The first set of the collected data can be categorised as internal. Internal data includes information from operating machines, personnel's data from the human resource department, financial data, research and developments, products' pricing, and the records of employees' timestamps. The second category of data collected can be identified as external data. The external data consists of information collected from websites, social media, market research, market trends, competitors, suppliers, and customers' feedback.

Finally, the study will closely examine how the collected data are stored before it is structured, processed, and analysed. Therefore, the concept of a technology called data lake, which is a repository for both structured and unstructured data will be introduced and discussed.

The recent revolution in information technology has enabled the digitalization and globalization in all sectors of business operations. Consequently, to be competitive and up to date, organizations are bombarded with enormous amount of data daily. Data are generated from different sources and different tools and mechanisms are put in place for their gathering.

The primary objective of the study is to find out if organizations pay any attention to the collected data. The study will also examine if there are continuous plans to maximize the importance of the collected data in order to make usage of it in daily business operations. Furthermore, the study will determine if the collected data influence decision-making processes thus, gaining and maintaining the competitive advantages.

The aim of the study is to discover if the respondent organizations know whether or not the massive amounts of collected data that are safely stored, prioritized, and utilized in business operations. Therefore, the aim is to determine if organizations are paying attention to the valuable gathered information to enable them to make informed business decisions.

Considering the width of the topic of data collection, storage and utilization, the focal point of the study is solely on the collection of data, the source provenance, the categorization of data, and its storage in raw format. The different technologies utilized to classify and process data are out of scope of this study. Consequently, the separate methods of extracting data for processing and analysing will not be discussed.

The study is conducted by sending a questionnaire survey of fifteen strict and unstructured questions to the Chief Executive Officers (CEOs) or managing directors of sixty-three organizations operating in different business sectors in Pietarsaari and Kokkola regions to respond. Upon reception of the responses, the answers will be sorted accordingly, the analysing process followed by the summarisation process will proceed, and finally, findings will be made.

2 DATA AND DATA COLLECTION

As an introduction to the topic in hand, I begin with a quote by Ratcliffe (2018), attributed to W. Edwards Deming: “In God we trust; all others must bring data.” Data in the sense of content and information can be compared to today's new gold. Whether organizations focus on data or not is the difference between whether they are in order or not because data today are at the centre of success of every business. Therefore, organizations are successful if their data are in good condition. Data are the capital that global organizations collect to be able to map and direct information and marketing to private individuals.

But also, for other organizations, it is important to keep track of their own information, both in terms of customers, staff, products, and services. Since data are important, there are threats connected to it, as a result data need to be handled efficiently and regularly. With data the goal is always to streamline and be able to guarantee the process quality. To be successful in business, organizations need to know what information the customers have already given to them to fulfil the needs and demands of the customers.

According to Cambridge Assessment International Education (2017), for many years, many philosophers held the belief that knowledge is power, and that knowledge comes from understanding of information. Information, in turn, is the assigning of meaning to data. Data in themselves are meaningless, however, information derives from processed data which illuminates insights for decision makers in all kinds of organizations.

In recent decades it has become evident that data are the most valuable asset any organization can have. Organizations thrive to gather as much data as possible in order to be able to make timely and informed decisions. Different mechanisms, technologies, and policies are adopted by various organizations to gain access to as much data as possible. According to Alghushairy & Ma (2019), the world is currently in the era of information science and new technologies, in which the use of information systems for information processing and storage has become a must have for organizations irrespective of the field of operation. Consequently, every modern organization needs reliable business information to allow effective business operations, to improve decision making process by gathering and evaluating data, to provide business operators with an accurate view of their activities, and, more widely, to adapt and be aware of trends and actual needs.

However, for mater of clarity, to be able to proceed and have a clear and common understanding, the word data need to be defined and explained. As a precaution, for the purpose of this thesis, I will not focus on data from the point view of computer scientists, but I will discuss data in the concept of source of business information that after being collected, processed, and analysed produce insights to business owners and decision makers, thus enabling informed decision-making processes.

2.1 Data and information

Data are the source of information. According to Cambridge Assessment International Education (2017), data can be described as a collection of text, numbers, videos, pictures, sounds, and the symbols without meaning. Data can also be the representation of facts, concepts, or instructions in a formalised manner suitable for communication, interpretation, or processing by people or by automatic means. (Checkland & Holwell 1998.) As defined by Zins (2005), data are everything or every unit that could increase the human knowledge or could allow to enlarge our field of scientific, theoretical, or practical knowledge, and that can be recorded, on whichever support, or orally handed. Data can incite information and knowledge in our mind.

Data have meaning beyond its use in computing applications oriented toward data processing. In fields such as electronic component interconnection and network communication, the term data are mostly separated from control information, control bits, and similar terms to distinguish the main content of a transmission unit. Moreover, in fields such as science, the term data are used to pinpoint a gathered body of specifics. That is also the case in fields such as manufacturing, engineering, finance, sales, marketing, and humanities. (Vaughan 2019.)

Information, on the other hand, is data that is organized and given the intended meaning for the recipient. Therefore, data are raw material that is altered into information by data processing. Thus, information can be defined in terms of its astonishing worth. It reveals to the recipient specifics that were unknown. (Wallace 2018.)

Data on themselves own are meaningless. Data only gets meaning and converts into information when it is interpreted. Data consist of raw facts and figures. When that data are processed into sets according to context, they supply information. Moreover, data refer to raw input that when processed or arranged makes meaningful output. Information is usually the processed result of data. When data are processed

into information, they become interpretable and become meaningful. (Cambridge Assessment International Education 2015.)

The Cambridge Assessment International Education (2015) develops this matter more by stating that in field of information technology (IT), symbols, characters, videos, audios, pictures, or numbers are considered data. Therefore, IT systems need these inputs of data to process in order to produce a meaningful interpretation. In other words, data in a meaningful form becomes information. Thus, information can be about facts, objects, concepts, or anything relevant to the topic concerned. It may further provide answers to questions like who, which, when, why, what, and how.

Consequently, as data and information have been fairly defined the next session will examine the collection of data, data provenance, and data characteristics.

2.2 Data Provenance

Data play a very important role in serving as the starting point whether it is in business, marketing, humanities, physical sciences, social sciences, or other fields of study or discipline. Data collection is the very first step in all processes that involve the usage of information and knowledge. (Kabir 2016a.)

The ultimate goal of data collection is to acquire all the necessary evidence that seeks to answer the questions that have been posed. Because of data collection, organizations can gather specific information that is helpful for decision making process. Accurate data collection is vital for any organization for maintaining the integrity of analysis, making informed business decisions, and protecting the quality of assurance. (Whitney, Lind & Wahl 1998.)

Additionally, Kabir (2016b) describes data collection as the process of gathering data and measuring information on variables of interest, in established systematic fashion that enables analysts and decision makers to answer queries, stated research questions, test assumptions, and evaluate results.

Data are collected from different sources for scientific research, business analysis, and many other purposes. As a result, the information technological developments in recent decades have increased the sources of data and this is demonstrated in the number of investments being made by organizations and governments in order to gather as much data as possible. Organizations are showing their continuous

commitment in the implementation of sophisticated, efficient, effective, and reliable technologies that are capable of collecting and processing data at very high speed. (Sivarajah, Kamal, Irani & Weerakkody 2017.)

In the past decades, the growth of the information systems, web applications and smartphones has led to significant rise in digital data generation for organizations in various fields of operation. Data now include pictures, texts, audios, and videos information, as well as system logs and websites and web applications activity records. Much of that is raw or unstructured data. (Sivarajah et al. 2017.)

Organizations collect data about the internal processes within the companies in order to determine areas of improvements by decreasing inefficiencies thus increasing profitability. Additionally, organizations collect data about their employees for further development of the business strategies and better planning. Furthermore, organizations are interested in the wellbeing of their employees, they plan and follow their career paths, expertise and skills and help them achieve their potentials.

According to Miloslavskaya & Tolstoy (2016), many technical tools contribute to data collection. Tools such as equipment's system logs are a great source of data in many business sectors. Machines' sensors, smart manufacturing machines applications collect considerable amounts of data at high velocity. Daily systems like human capital management software, various information systems such as ERP (Enterprise Resources Planning) systems, CRM (Customer Relationship Management) systems, financial reports, and sales reports enable organizations to gather data. Organizations collect data from various IoT (internet of things) devices, manufacturing equipment provide needed data, and many other systems are deployed in order to gather data for gaining insights.

However, it is worth mentioning that all data are not only collected via digital means. Interpersonal collaborations also help organizations in collecting valuable data. Employees express their problems, as well as improvement suggestions, that can be helpful for organizations. Collaboration with different business partners is a valuable source of data for many organizations because the people on the field have insights that can prove beneficial if taken into consideration. Furthermore, organizations are interested in following trends and market research, and are keen in collecting important data from websites, and feedback from their business partners.

2.3 Data categorization

Considering the purpose and manner on which data are being gathered, data can be categorized in several ways. Data can be considered primary data, meaning that they are gathered directly from the source. Secondary data refer to the data that has been gathered and analysed by a third party.

Moreover, data can be considered qualitative or quantitative. Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code. Quantitative data, on the other hand, are measures of values or counts and are expressed as numbers. Data collected about a categorical variable will always be qualitative and data collected about a numeric variable will be quantitative. (Australian Bureau of Statistics 2020.)

Despite the characteristics of data mentioned above, data can be considered raw (unprocessed) or processed and analysed if it has been manipulated by machines or humans. Additionally, data can be structured or unstructured. (Praveen & Chandra 2017.)

However, for the sake of this thesis I will focus on the structure of data and the area of data provenance. I will consider data as being internal or external based on the provenance of collection. Figure 1 demonstrates different types of data discussed below.

2.3.1 Unstructured, structured data and metadata

Data, whether structured or unstructured, are the life force of any organization and should be at the heart of every decision-making process. Organizations need data to be aware of what is happening within and outside their business peripheries. This data can be structured signifying classified and organized, or unstructured meaning unclassified and unorganized. (Praveen & Chandra 2017.)

According to Praveen & Chandra (2017) all data are first gathered as raw in its state before being sorted, processed, and analysed for different usage. Typically, the data that are most readily available for use are not in a state in which they can be used easily. Such data are difficult to manipulate and typically need to be processed in some fashion before they can be used. Data which have yet to be processed are sometimes referred to as raw data, although source data are a more useful term. These data, being in their original state, are unstructured and have not yet been structured in a predefined manner.

Unstructured data are typically text-heavy, like system logs, sensors' generated data, and handheld devices' data. They can also include images, videos, and audio.

Unstructured information is growing quickly due to increased use of digital applications and services. Unstructured data are valuable to businesses if analysed and interpreted correctly. It can provide a wealth of insights that statistics and numbers just cannot detail. Additionally, unstructured data cannot be easily stored in a traditional column-row database or spreadsheet like a Microsoft Excel table. It is therefore more difficult to analyse and not easily searchable. (Grossman 2019.)

Currently, more than 80% of the data on the Internet are unstructured data (Allahyari, Pouriyeh, Assefi, Safaei, Trippe, Gutierrez & Kochut 2017). Unstructured data usually refers to information that does not reside in a relational database. In other words, the data structure of unstructured data are irregular or incomplete and there is no predefined data model. It should be noted that although some document formats like comma separated value (CSV), JavaScript object notation (JSON), and extensible markup language (XML) have some organizational properties, they usually do not have a clear predefined data model. Compared to structured data, these data are still difficult to retrieve, analyse and store. Unstructured data are easily processed by humans but are very hard for machines to understand. (Allahyari et al. 2017.)

There are some characteristics that define the unstructured data. Unstructured data have no internal identifier to let search functions recognize it. Unstructured data do not follow any semantic or rules, they are gathered in the state that they are generated. Unstructured data lack any particular format or sequence and do not possess easily identifiable structure. Therefore, due to their lack of identifiable structure, they cannot be used by computer programs easily. These data are generated by various sources in different formats and need some work to them before they can be used by organizations. (Praveen & Chandra 2017.)

As discussed earlier, each data have a source provenance. Web pages produce enormous amounts of unstructured data through inbuild data collecting applications. These enormous data can be in form of images (JPEG, GIF, PNG, etc.), videos, sounds, and text files. Unstructured data can be in forms of memorandums, various reports and systems logs, text documents and so on. Research surveys, and customers' feedback can generate a considerable amount of unstructured data that need sorting before processing and analysing. (Allahyari et al. 2017.)

There are many advantages the unstructured data bring to organizations. Because of their lack of proper format and sequence, the data are not constrained by a fixed schema. Moreover, the gathering is very flexible due to absence of predefined schema. (Praveen & Chandra 2017.)

Regarding the unstructured data concept, unstructured data are mobile and portable. Because it is generated and stored efficiently, unstructured data prove to be flexible. Unstructured data enriche organizations' data and enable decision-makers to work effectively and proactively. Moreover, unstructured data are scalable, their volume increase at a very high speed, and they can deal easily with the heterogeneity of provenances. These types of data have a variety of business intelligence and analytics applications. (Sivarajah, Kamal, Irani & Weerakkody 2017.)

However, as with any solution, Sivarajah et al. (2017) state that there are some disadvantages with unstructured data. It is difficult to store and manage unstructured data due to lack of schema and structure. Therefore, indexing the data is difficult and error prone due to unclear structure and lack of predefined attributes. The data are disorganized, and it consumes enormous time and resources to extract insights from them. Due to their disorganization, search results may not be very accurate if extra precautions are not taken. Since data originate from various sources, ensuring security to data is a difficult task.

In contrast to unstructured data, structured data are the data which conform to a data model. They have a well define structure, they follows a consistent order, and can be easily accessed and used by a person or a computer program. Structured data have a well-defined structure that helps in easy storage and accessibility. Data can be indexed based on text string as well as attributes. This makes search operation easy and quick. Structured data are usually stored in well-defined schemas such as databases. It is generally tabular with column and rows that clearly define its attributes. Structured data exist in a format created to be captured, stored, organized, and analysed. They are neatly organized for easy access. Consequently, structured data bring inherent benefits when dealing with high volumes of information. (Praveen & Chandra 2017.)

However, according to the article "The age of analytics, competing in a data driven world" published in 2016 by the McKinsey Global Institute, it is important to mention that despite the amount of data collected, structured data account for only about 20% of data within organizations.

Structured data depend on the existence of a data model – a model of how data can be stored, processed, and accessed. Because of a data model, each field is distinct and can be accessed separately or jointly along with data from other domains. This makes structured data extremely powerful. It is possible to quickly aggregate data from various locations in the database. However, structured data records can hold unstructured data within it. In a database, data are well organised so that data definition, format and meaning is explicitly known. (Gartner 2008.)

Structured data reside in fixed fields within a record or file. Similar entities are grouped together to form relations or classes, and entities in the same group have same attributes thus, enabling easy to access and query. As a result, data elements are addressable, efficient to process and analyse. (Tandy, Ceolin & Stephan 2016.)

Since all data have a provenance, structured data are not an exception. Structured data come from structured query language (SQL) databases, spreadsheets such as excel, smart manufacturing systems, online forms, sensors such as global positioning system (GPS) or radio frequency identification (RFID) tags, systems' logs and journals, handheld devices, retail, and ecommerce. Considering the nature of structured data, there are advantages that cannot be ignored. Structured data make operations, such as updating and deleting, easy due to its well-structured form. Data can be captured and stored securely, and the scalability is not a problem in case there is incrementation of data. Data analysis is quicker because data is already structured and ready for processing. (Gartner 2008.)

Nowadays, organizations possess both unstructured and structured data, and the provenance varies from machine generated to human generated data. Both these data are vital when used in daily business decision making processes.

According to Daniel & Daniel (2012), metadata is data about data. The prefix meta in English is used to express the idea that some information is about its own category. Hence the meaning of metadata is data about data. Metadata can be found inside a file where an ordinary computer user will not see it, or in an external data store such as internet history files. The information stored within metadata can be used to build timelines, establish alibis, and so much more. Metadata can shed light on a particular issue for example in an investigative case.

Metadata is data, which means that it can be modelled and managed the same way other data are handled. There are numerous types of metadata that are more interesting and valuable because they convey information about the utility of the data as well as facilitating methods for reusing or repurposing information in different ways. The tools for handling metadata still require some degree or maturation for them to reach the point where they can satisfy the collected needs of the organization. (Loshin 2013.)

The metadata is useful in different ways within organizations. There are many sectors in the organization that make use of the metadata. Some of the prominent users of metadata include the system developers, the data administration professionals, and the various end users. The system developers utilise metadata to determine how new systems must be interconnected and made compatible with older existing systems. Likewise, the data administration professionals use metadata to determine how the data model needs to be changed, or if there is a new data model, how this will fit with the existing data model. Finally, the various end users utilise metadata to help in the creation of queries. (Inmon & Lindstedt 2015.)

In the context of a file system, metadata is additional information about content in blocks. In the context of file analysis, metadata is information stored within the file itself that provide some possibly interesting but otherwise nonessential information about the file. Metadata is included to provide context or extended information that is outside of the scope of data themselves. The value of metadata is highly dependent on the nature of a given examination and the types of files being examined. (Altheide & Carvey 2011.)

Metadata plays a very different role than other data contained in a data warehouse and is important for many reasons. For example, metadata is used as a directory to help the decision support system analyst locating the contents of the data warehouse. It is also used as a guide for data mapping when data are transformed from the operational environment to the data warehouse environment. Metadata should be stored and managed persistently. (Han & Pei 2012.)

2.3.2 Internal data

Internal data is data retrieved from inside the organization to make decisions for successful operations (Arthur 2013). These data are important to determine whether the strategies the organization is currently using are successful or if modifications are necessary. As the word internal indicates, data are internal if an organization generates, owns, and controls them. There are many benefits to why organizations are interested in the internal data. Internal data can benefit organizations that want to improve efficiency and productivity and organizations that are struggling to be profitable. (Baud, Franchot & Roncalli 2002.)

Internal data are the engine that helps decision makers run and optimize their daily business operations. These data are reliable because the sources of provenance are accurate, thus verification is not always necessary. While internal data are data generated from within the organization, they cover areas such as operations, maintenance, personnel (human resources), sales, marketing, and finance. Each area provides a unique perspective, yet the data connect the departments. Furthermore, internal data are extremely important for the success of any business because they are readily available for process and analysis whenever there is such a need. The ability to make quick decisions is enhanced by rapid access to data. Another advantage of internal data is that they show a very clear trajectory of the organization without any dependency on outside resources. (Arthur 2013.)

2.3.3 External data

External data are the data generated outside the organization and therefore, the organization neither owns nor controls them. These data can range from economic trends, market research, consumers or customers’ behaviour and feedback, and government regulations within an industry. (Schatsky, Camhi & Muraskin 2019.)

Different tools and methods such as website data collectors, line of business applications, questionnaire surveys, various market research are utilized to collect data from outside sources. External data help organizations' decision makers better understand their customer base and the competitive landscape. Organizations need a clear view of what is happening outside to make truly insightful business decisions.

According to Schatsky et al. (2019), organizations are increasingly seeking better insights by collecting and analysing third-party data. This is not an exception for organizations in industries including financial services, logistics, technology, health care, retail, and the public sector. Most organizations are using external data to gain new insights that can help increase efficiency and revenue.

Mining external data for insights is increasingly important, therefore, organizations know they can gain valuable insights by analysing the data they generate from their operations. However, because internally generated information can leave gaps and show a one-sided picture of operations, organizations are increasingly moving to incorporate new, non-traditional, and external sources of data into their analyses. These data can include almost anything.

Collecting and analysing external data enable organizations to foresee and prepare for risks and opportunities which they could otherwise easily have missed with inputs limited to data generated from internal operations, customers, and first-tier suppliers. With globalisation organisations increasingly operate as part of networks consisting of business partners such as suppliers, resellers, channel partners, regulators, and other stakeholders. These networks are often globally distributed and potentially affected by financial, political, and/or environmental factors. (Arthur 2013.)

Furthermore, collecting and analysing external data illuminate how factors such as shifting market trends and behaviours, competitor initiatives, or geopolitical events can affect their business. External data sources are helping organizations to personalize marketing offers, improve human capital decisions, gain new revenue streams by launching new products or services, enhance risk visibility and mitigation, and better anticipate shifts in demand for their products and services.

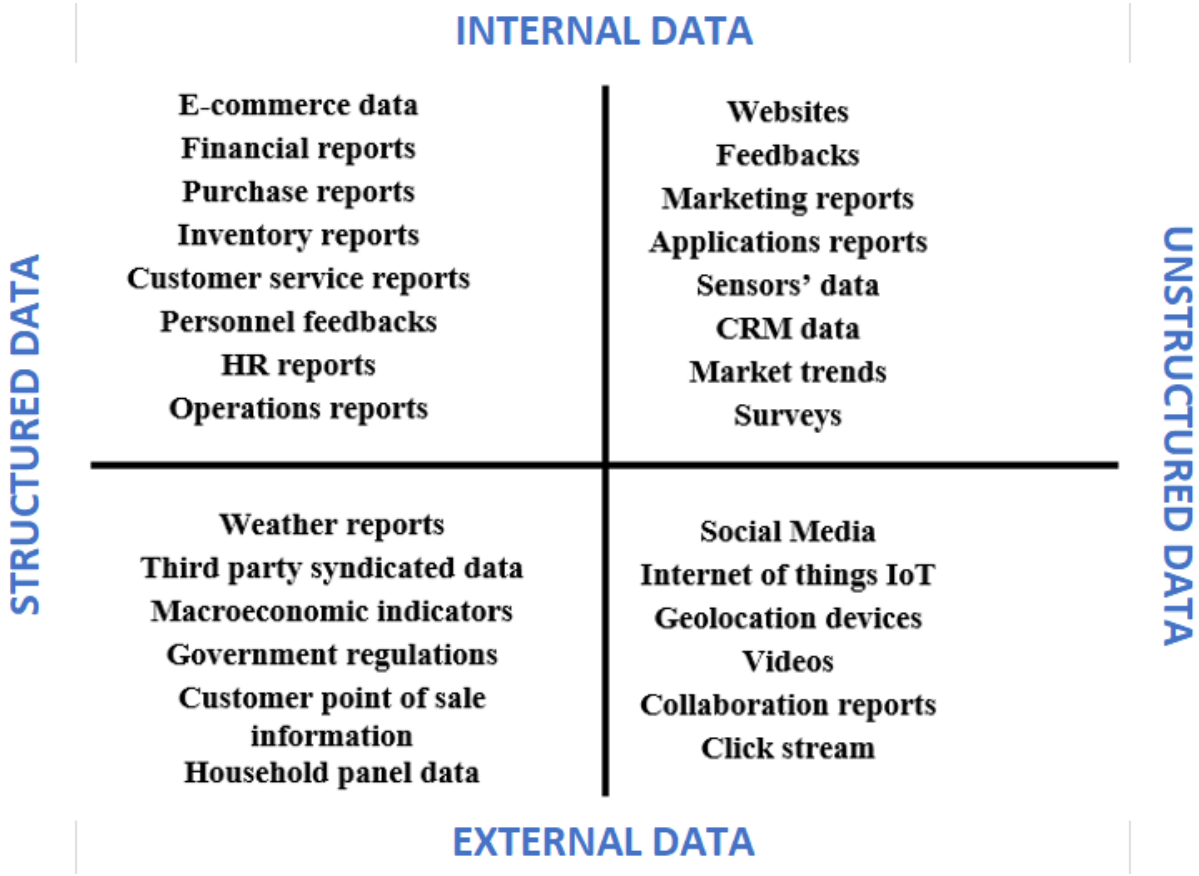


FIGURE 1. Different types of data (adapted from Wilson 2019)

3 DATA STORAGE

A continuous phenomenon of our generation is the industrialization and the recent digitalization trends which have resulted in the exponential increase in volume of data being generated in organizations quotidianly. According to Alghushairy & Ma (2019) the data are generated by humans and machines irrespective of the industry or geolocations. Therefore, as the value of generated data increases and organizations become more dependent on the insight gained from data, reliable means of storing and archiving the data is an integral process for any operating organization.

Organizations invest enormous amounts of money and time to gather data for analysis and gathering of insights for better decision-making processes. As data are produced, the demand for safe and accessible storage increases as well. In this section I will define the term storage and its purpose. I will discuss different types of storage. Finally, characteristics of storage will be introduced.

The world is currently in the era of information science and new technologies, in which the use of information systems for information processing and storage has become indispensable for organizations irrespective of the field of operation. Consequently, every modern organization needs reliable business information to allow effective business operations, to improve decision making process by gathering and evaluating data, to provide business operators with an accurate view of their activities, and, more widely, to adapt and be aware of trends and actual needs. (Alghushairy & Ma 2019.)

Alghushairy & Ma (2019) define data storage as the process of storing and archiving data in electronic storage devices that are designed and dedicated for conservation, in return, making data accessibility a priority at any time. Therefore, conventional storage devices are hardware that are used for reading and writing data through a storage medium. Storage media are physical materials for storing and retrieving data.

According to an article in Techopedia website (2012), the most common data storage devices are hard drive disks, flash drives, and cloud storage while compact disks (CDs) and digital video disks DVDs) are becoming obsolete to organizations due to the incapacity in coping with the amounts and speed at which data are collected and stored.

Recently, the term big data has been used interchangeably to describe the amounts of data being generated and stored by organizations globally. The term big data reflects not only the massive volume of data but also the increased velocity and variety in data generation and collection, for example, the massive amounts of digital photos shared on the internet, social media networks, and even the Web search records. Moreover, many materials such as eBooks, online-newspapers, and blogs can also be data sources in the digital world. (Eiras 2011.)

Since data are the all-important asset for any organization, it is necessary to store data in appropriate ways to support data discovery, access, and analytics. To match the demands and challenges posed by data storage, various data storage devices and technologies have been developed to increase the efficiency in data management and to enable information extraction and knowledge discovery from data.

Eiras (2011) insists that in domain-specific fields, a lot of scientific research has been conducted to tackle the specific requirements concerning data collection, data formatting, and data storage, which have also generated beneficial feedback to computer science. This fact has encouraged world researchers to focus on new techniques to design devices with larger storage capacity, as it is the case of cloud storage solutions, and high-capacity storage devices. These devices can store the entire volume of the informational material, thereby increasing the storage capacity in comparison with two-dimensional devices that only store the information on the surface. (Eiras 2011.)

In conclusion, data storage is a key step in the data science life cycle. At the early stage of the cycle, well organized data will provide strong support to the research program where the data are collected. At the late stage, the data can be shared and made persistently reusable to other users.

3.1 Types of storage

Bespalov, Michel & Steckler (2020), specifically discuss different types of storage. They focus on the actual medium and employed technology. I will consider the most prevalent options such as magnetic and optical technology, and semiconductor technology.

In the magnetic technology data are stored using the magnetization patterns on a special surface. Hard disk drives (HDD), and magnetic tapes are considered magnetic technology. Within the optical technology (compact disks and Blu-ray), data are stored in deformities on a circular surface which can be

read when being illuminated by a laser diode. Regarding the semiconductor technology, data are stored using semiconductor-based integrated circuits. Traditionally, the semiconductor technology was used for volatile storage in which data are lost if electric power is not supplied. In contrary to magnetic or optical storage, solid-state drives (SSD) are now included in consumer computers, offering a non-volatile option with superior access speeds to their magnetic counterparts. When dealing with these options, various aspects such as convenience, costs, and reliability, need to be taken into consideration. (Blumzon & Pănescu 2019.)

3.2 Characteristics of storage systems

Technology for storing data can be of less relevance but other characteristics can play an important role when choosing a solution. The location of the data storage is the first considered aspect.
Storing data on a local machine is advantageous as it allows users to work offline but might place obstacles in organizations when attempting to share it with a larger team. This requires that the owner of the machine is fully responsible for data preservation in case of hardware failure. Moreover, the data being processed by the user might prove to be outdated and duplicated. However, storage facilities managed at the institutional level, such as storage area network (SAN) systems, move the burden of managing data storage from the individual user to specialised personnel, providing higher reliability and enhanced possibilities for sharing data among peers within and outside organizations. Finally, data can be stored offsite in specialised facilities. This model became prominent with the advent of cloud systems, such as Amazon Web Services, Microsoft Azure or Google Cloud Platform, and has benefits in terms of reliability, scalability, and accessibility. (Blumzon & Pănescu 2019.)

According to Blumzon & Pănescu (2019), cloud storage might be preferred when the individual user in the organization does not possess the required resources for managing a storage solution, when large quantities of data need to be stored, or when data need to be shared across a large network of collaborators.

Since data storage is vital for organizations, storage solutions must be redundant. No storage system is perfect and therefore it is important that data are copied and stored on different systems simultaneously. The higher the number of copies and the broader their distribution, the higher the guarantee for their persistence is. Simply having data stored on a medium does not provide guarantees that, over time, it would not become inaccessible. Tape drives and hard disks can become demagnetised, hence

corrupting the stored data. Furthermore, as technology evolves, so do the methods for storing data. Storage and archival systems need to account for this and migrate data from old devices, such as floppy disks and hard disks, to modern technological requirements while keeping the integrity of the data. (Blumzon & Pănescu 2019.)

3.3 Cloud storage

Recently the cloud has become the ultimate data repository. It is cheap, easy to use, and fast. There are enough resources that are self-repairing, and data can be recovered from anywhere at any time. Cloud storage is a system that provides functions such as data storage and business access. It assembles a large number of different types of storage devices through the application software which are based on the functions of the cluster applications, grid techniques, distributed file systems, etc. Cloud storage can simply be understood as large capacity storage in cloud computing. Cloud storage system architecture mainly includes storage layer, basic management layer, application interface layer and access layer. Cloud storage has enormous potential and will continue to be part of organizations even in the future. (Liu & Dong 2012.)

One of the core concepts of cloud computing is reducing the processing burden on user's terminals through continuously enhancing the clouds' handling capacity. Eventually user's terminals are simplified into simple input and output devices. Users can use the powerful computing and processing function on clouds, and they can order their service from the cloud according to their own needs. (Liu & Dong 2012.)

3.4 Data repositories

A data repository is a tool that is common in scientific research but also useful for managing business data. Data repositories are the primary conduits by which researchers and business analysts store and share their data and software packages. (Johnston 2017.)

A data repository refers to an enterprise data storage entity (or sometimes entities) into which data have been specifically partitioned for an analytical or reporting purpose. The data repository is a large

database infrastructure, several databases, that collect, manage, and store data sets for data analysis, sharing and reporting. (Qamar 2015.)

Data have established themselves as the most important asset to organizations for decisions making processes. Therefore, tools that can collect, store, and help in analysing data are continuously required.

In recent years, data repositories have been rapidly developed, and many public repositories such as UCI, GitHub, Zenodo, and Dryad have been in use in various domains. These repositories provide a large volume of high-quality data for research, business analysis and practice. (Xie, Wang, Kim, Lee & Song 2021.)

Many terms such as a data library or data archive can be interchangeably used for a data repository. Thus, data repository is a general term referring to a data set isolated to be mined for data reporting and analysis. Moreover, data completeness, accessibility, ease of operation, and credibility positively affect the satisfaction of data users. (Faniel, Kriesberg & Yakel 2015.)

The purpose of a data repository is to keep a certain population of data isolated so that it can be mined for greater insight or business intelligence and to be used for a specific reporting need. Due to the value of the stored and analysed data, organizations can make decisions based upon more than presumptions and instincts. However, using data repositories as part of data management is another level of investment that can improve business decisions. Isolated data, where the data are clustered together, allows for easier and faster data reporting and analysing. When data are isolated, compartmentalized, and preserved, database administrators can effectively track problems. (Qamar 2015.)

However, as with many technologies, there are several vulnerabilities that exist in data repositories that enterprises must manage effectively to mitigate potential data security risks. Large amount of data makes systems slow. Therefore, there must be a balance between database management systems and data growth. Since an impaired system could affect all the data it is imperative to back up the databases and isolate access applications. Furthermore, unauthorized users can access all sensitive data more easily than if they were distributed across several locations. (Kleppmann 2017.)

For the sake of this thesis, I will discuss different types of repositories. I will introduce data repositories technologies such as data lake, data warehouse, data marts, metadata repository and data cubes. However, the data lake repository technology will be the main focus of the thesis.

3.5 Data Lake

As defined by Ziegler, Reimann, Keller & Mitschang (2020), a data lake is an underlying storage component that handles huge volumes of heterogeneous unstructured and structured data.

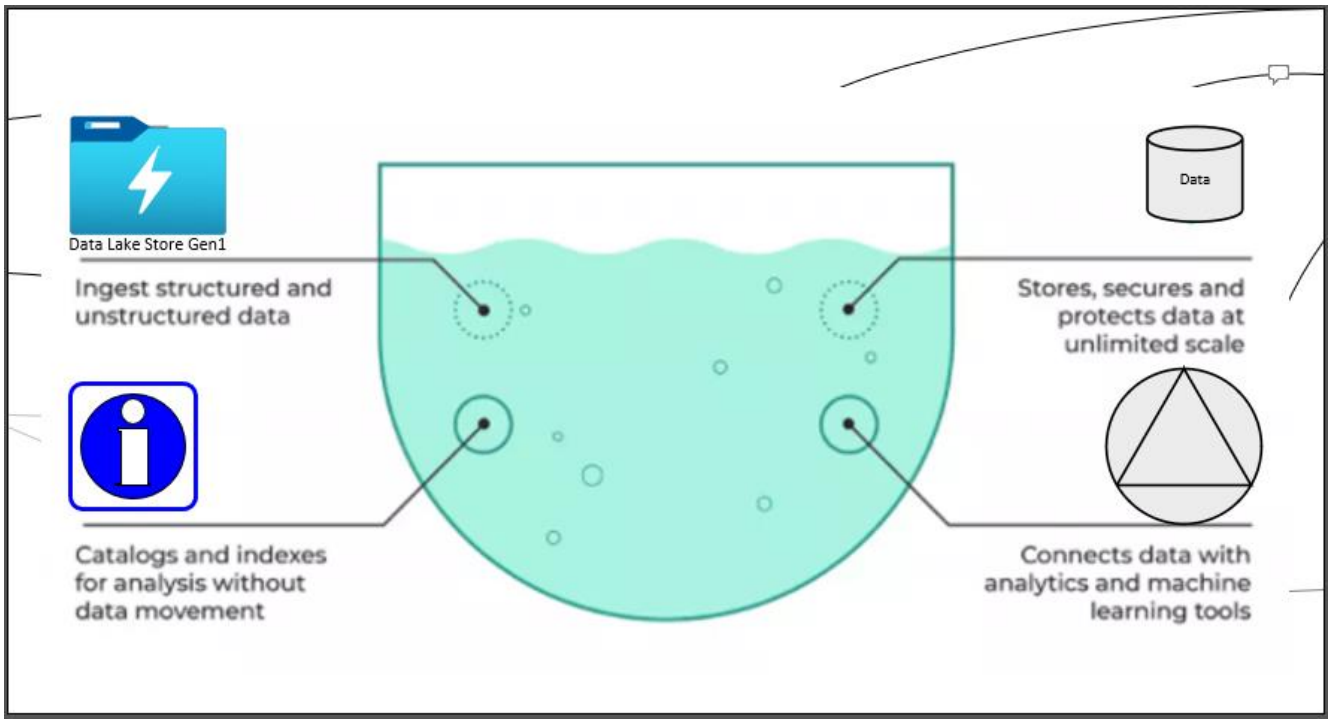


FIGURE 2. Data lake features (Singman 2021)

As described in figure 2, a data lake is a storage for any kind of data in high volumes that are retrieved from multiple sources. Consequently, both unstructured and structured data are stored in the data lake in their unprocessed raw data format (Ziegler et al. 2020.)

Moreover, the data lake is responsible to supply data to other departments for further use, such as for decision making processes, or for data analysis purposes. A data lake is a single store of all enterprise data including raw copies of source system data and transformed data used for tasks such as reporting, visualization, advanced analytics, and machine learning. (Ziegler et al. 2020.)

Grossman (2019), states that sometimes the term ‘data lake’ is used when data are stored simply with digital IDs and metadata (shallow indexing), but without a data model. Therefore, data models and schemas are used when the data are written or when the data are analysed not when the data are stored.

According to Llave (2018), the data lake approach has emerged as a promising way to handle large volumes of structured and unstructured data. This data collection and data analysis technologies enable organizations to profoundly improve their operative processes and decision-making abilities.

Modern technologies like data lakes have made it possible to acquire data without a full understanding of the data's structure. Also, the data lake technology has emerged as new type of data repositories that enables storage and processing power to support the analysis of large unstructured data sets.

The term data lake was invented by James Dixon, the chief technology officer (CTO) of Pentaho, a business intelligent software owned by Hitachi Vantara corporation. His aim was to convey the concept of a centralized repository containing virtually inexhaustible amounts of raw data for analysis or undetermined future use. As a result, organizations across various industries are beginning to place their data into data lakes without performing any data transformations. (Llave 2018.)

Based on a study conducted by Llave (2018) on the data lake topic, the data lake can be defined from two different perspectives: a technology perspective, and a business perspective. From the technology perspective, a data lake is the collection of technologies with data that one needs to store in some specific format. So, a data lake is not one data lake; it is many technologies that serve the data's need. A data lake is a central repository of any type of data and a central repository of truth. However, from a business perspective, a data lake is a capability of the business where one can get raw data emerging from different source systems. A data lake is the place where all data in the enterprise can be collected.

Figure 3 describes the purpose that characterises a data lake in different organizations. Data come in from both internal and external sources. Both unstructured and structured data can be stored in a data lake before being transferred to a data warehouse for queries and processing or for direct analysis.

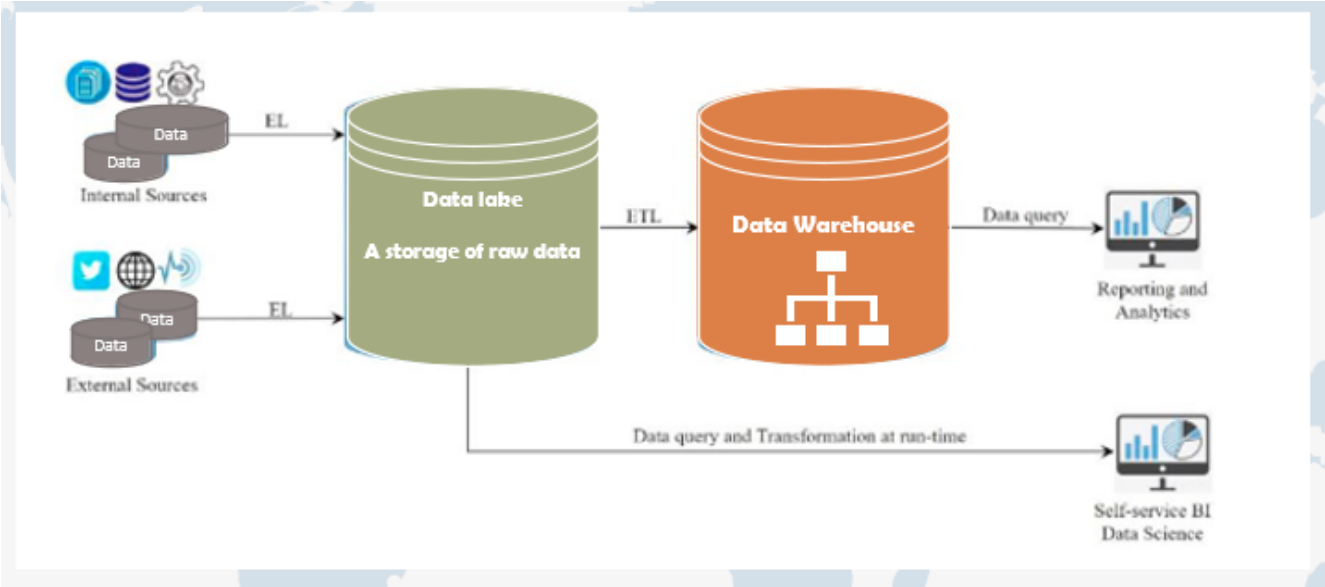


FIGURE 3. Different purposes of data lakes (Llave 2018)

As illustrated in figure 3, a data lake is firstly a staging areas or sources for data warehouses. A staging area is a temporary location between a data provenance and a data warehouse. The staging area is a storage area, typically a relational database, to temporarily keep a copy of the source data as a step on the way to the data warehouse. The main purpose of the staging area is to avoid heavy processing and potential overload of the source system that might be critical for businesses when transforming the data on the way to the data warehouse. A data lake is a storage area that keeps a permanent copy of different types of source data, both structured and unstructured. In a data lake data can both be kept for current defined needs and for future undefined needs. Secondly a data lake can be used for storing histories or archiving. A data lake can also be used for offloading archived data from data warehouses. Therefore, a data lake is a useful component in any data warehouse architecture and that it can be seen as an extension of the concept of business intelligence and for data science and advanced analytics. (Grossman 2019.)

The data in the data lake are stored as they are extracted, on the same data structure as in the source system or as received, without any transformations. With the introduction of the Internet of Things and sensors, organizations need some place to store all these various data that come from new technology. To be able to store these data, relational databases, like SQL, would not be fit for this purpose. Therefore, the data lake is the ultimate solution, hence, the sole purpose of the data lake is to store the unstructured data or the odd data that come from middle things, like sensor devices and web logs. (Llave 2018.)

However, according to Llave (2018), data scientists and business analysts are the frequent users of data lakes. Data lake technology is useful for exploration and advanced analytics. This means that analytics can directly be done in the data lake, and when good data are found, they then can be moved into the data warehouse for further analyses and reporting.

The organizations can use data in the data lake for research and development, and experimentation to be acquainted with data for query from the data warehouse. Moreover, data lake can be used as direct sources for self-service business intelligence for reporting and analytics tools (FIGURE 3.) Organizations can produce new reports built on the information found in the data lake. Consequently, organizations can also use self-service business intelligence directly on the data lake in accordance with the data warehouse. A semantic layer is applied in between the data lake and self-service business intelligence tools. (Llave 2018.)

However, as with all technologies, the research conducted by Llave (2018) revealed several challenges related to data lake, for example challenges related to data stewardship. Good data stewardship involves good care and secure measures. Another example of challenge is data governance. Organizations that need to secure and obscure confidential data may struggle to implement this in a data lake. Furthermore, there are challenges concerning the analytical process. In the data lake, data are in the original format resulting in a higher requirement for expertise and excellence regarding how to select, analyse, and interpret the data. Also, when it comes to data quality, precaution must be taken when dealing with devices providing data, making sure that all the sources are functional and reliable.

Finally, data retrieval poses another challenge related to data lakes. The difference between a data lake and a data warehouse is that, in a data warehouse, data are transformed before they are stored in the data warehouse while in a data lake, data are stored in the original format. To create insight, the transformation is done afterwards. Organizations must therefore extract the needed data and build a program to cleanse it for intended purpose. (Llave 2018.) According to Grossman (2019) data lake technologies involve less effort during data gathering, but more effort during data retrieval.

3.6 Data warehouse

According to Tupper (2011) data warehousing is currently one of the most important applications of database technology and practice. A significant proportion of IT budgets in most organizations may be devoted to data warehousing applications.

A data warehouse is a central repository for the important data of an organization. A data warehouse is a core enterprise software component required by a range of applications related to the business intelligence. Data from multiple operational systems are uploaded and used for predictive analysis and for the detection of hidden patterns in the data. Data models developed using statistical learning theory allow organizations to optimize their operations and maximize their profits. (Marinescu 2018.)

Data warehouses can be very powerful and useful solutions for an organization to use in data consolidation and reporting. However, it tends to take a very long time to add a new data source to a data warehouse, from concept to implementation. Data warehouses are an integrated hub of data that can be used to support business analysis and reporting. Many organizations implement multiple data warehouses to support different geolocations or functions within the organization. (Reeve 2013a.)

All the data consumers who want to access specific data can get it from a single place rather than having to go to various operational applications independently. Data warehouses and business intelligence tools make more and better data available as the number of end users increases. The demand for better performance is more important than ever. In addition, the types of queries are increasing in complexity. Data warehouses are being used to support new types of e-business initiatives including customer relationship management (CRM) and supply chain management. CRM helps attract new customers and develop customer loyalty and is important in the retention of existing customers. (Hobbs & Smith 2005.)

A data warehouse contains the information about organizations' customers and is often the integration point for sales, marketing, and customer care applications. Thus, ensuring the availability of the data warehouse is becoming more and more mission critical for many businesses. As data warehouses are becoming more operational in nature, feeding information back to the OLTP (online transactional processing) systems, users need access to the data warehouse without any system downtimes. This is specifically important for organizations that operate globally. (Hobbs & Smith 2005.)

3.7 Metadata repository

Metadata repository is a common term for a computerized database containing metadata to support the development, maintenance, and operations of a major portion of an organization's systems. Among other things, such a repository can be the foundation for a data warehouse. The first metadata repositories were the data dictionaries and copy libraries that accompanied programs in the 1970s and 1980s. A data dictionary was simply a listing of the fields contained in a record of a particular type in the files of a traditional mainframe data processing application. Sometimes this was accompanied by definitions of the meanings of each file and field. (Hay 2006.)

A metadata repository is a tool for storing metadata. In some cases, the metadata repository is largely focused on recording physical and technical metadata, such as data models, database structures, metadata associated with Business Intelligence (BI) tools and file structures. Additionally, the metadata repository can provide impact analysis for proposed changes, and lineage (where data came from and how it was manipulated). (Plotkin 2021.)

According to Han & Pei (2012), a metadata repository should contain a description of the data warehouse structure, which includes the warehouse schema, view, dimensions, hierarchies, and derived data definitions, as well as data mart locations and contents. Operational metadata, which include data lineage (history of migrated data and the sequence of transformations applied to them), currency of data (active, archived, or purged), and monitoring information (warehouse usage statistics, error reports, and audit trails).

Finally, a metadata repository links the metadata from the various tools and engines and provides an enterprise view and audit trail of the movement and transformation of data around the organization. (Reeve 2013b.)

3.8 Data marts

A data mart is a subject-oriented data repository, similar in structure to the enterprise data warehouse, but it holds the data necessary for the decision support and BI needs of a specific department or group within the organization. A data mart could be constructed solely for the analytical purposes of the specific group or could be derived from an existing data warehouse. Data marts are also built using the star join structure. It is worth mentioning that there are differences between a data mart and a data warehouse, mostly due to the different natures of the desired results. (Loshin 2013.)

Data warehouses are meant for more loosely structured, exploratory analysis, whereas data marts are for more formalized reporting and for directed drill-down. Because data marts are centred on specific goals and decision support needs of a certain department within the company, the amount of data is much smaller but focused on data of the particular department's operation. This implies that different departments with different analytical or reporting needs may need different kinds of data mart structure. A data mart is likely to be configured for more generalized reporting for the specific business users within the department. Standard reports are more likely to be generated from the data mart, which will be much smaller than the data warehouse and will provide better performance. (Loshin 2013.)

3.9 Data cubes

Data cubes allow data to be modelled and viewed in multiple dimensions within data warehouses. Data cubes are defined by dimensions and facts. Generally, dimensions are the perspectives or entities with respect to which an organization wants to keep records. These dimensions allow the organization to keep track of its activities and operations. Each dimension may have a table associated with it called a dimension table which further describes the dimension. A dimension table can be specified by users or by experts, or automatically generated and regulated based on data distributions. The data cubes are a metaphor for a multidimensional data storage. (Han & Kamber 2006.)

A data warehouse is usually modelled by a multidimensional data structure, called a data cube, in which each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure such as count or sum (sales amount). As a data cube provides a multidimensional view of data it allows the precomputation and fast access of summarized data. Data cubes facilitate the answering of queries as they allow the computation of aggregate data at

multiple granularity levels. Traditional data cubes are typically constructed on commonly used dimensions using simple measures. (Han & Pei 2012.)

4 RESEARCH ANALYSIS

The aim of this chapter is to analyse the results of the findings of the data from the questionnaire survey conducted by the researcher. The gathered data are analysed according to the objectives of the research. The data were collected by sending an electronic questionnaire survey which respondents filled and submitted. The survey was conducted via the online tool called Webropol surveys service.

4.1 Key informants

As stated by Wellington & Szczerbinski (2007), a key informant is a person who is a key figure in a piece of qualitative research. It is essential to identify the key informants in any research and they can range from one person to several persons depending on the case study or focus group of the research.

Therefore, for the purpose of this research, the key informants are the Chief Executive Officers (CEO), or the Managing Directors of different organizations in Pietarsaari and Kokkola regions in Finland.

As earlier mentioned in the introductory session of this thesis, in order to find information about the importance of data within organizations in our region, it is vital to directly interview the strategists and decisions makers. If decision makers value the insight gained from gathered data, they will invest and put in place all the possible measures, policies, and technologies to collect, store, process and analyse data. The gathered and analysed data will enable them to gain the valuable insights leading to better planning and informed decisions making processes for gaining and maintaining competitive advantages.

4.2 Research questions

The strict and unstructured questionnaire survey is purely based on the organizations. The focus is the collection of data and the storage of data for later usage. The research used six themes presented in (TABLE 1) to examine the impact and value of data in business operations. To achieve the objectives,

the themes were dealing with sectors of business operation, data collection in general, internal data collection, external data collection, data storage, data dependency and commitment to data gathering in the future.

TABLE 1. Research question themes

Information about the organization
Information about data collection
Information about internal data collection and sources
Information about external data collection and provenance
Information about data storage
Information about organizations' dependency on collected data

TABLE 2. Research questions

In which business sector does your company operate?
Numbers of employees
What is your main market?
Does your company collect information about the employees (career paths, expertise, skills, and needs for further training)?
Does your company collect information from equipment's system logs?
What are the tools, sources used to collect information?
From which of the following does your company collect data for day-to-day business operations?
Does your company collect data from external sources to make business decisions?
Which of the following external sources does your company use to collect data?
Which of the following describes the data storage in your company?
Do the insight gained from collected data assist your company in decision making process?
How dependent is your business on the collected data?
Do you see advantages in making informed decisions based on gained insights?
Is your company committed to continue collecting data to gain and maintain competitive advantage?
Do you have anything to add or comment on this topic?

Table 2 provides the detailed questions that were sent to organizations in order to collect vital information for the purpose of the thesis.

The questions asked through the survey endeavoured to collect information from the CEOs about their organizations, about data collection policies, and about the provenance of data. The questions also focused on the tools used to collect data, the storage policies, and I tried to find out if the organizations are dependent on collected data for decision making processes. Finally, the questions were intended to discover if the organizations are dedicated to continuing gathering data in the future.

The questionnaire survey was sent to CEOs of 63 different organizations. The reason for sending directly to CEOs was based on the purpose and the impact of the research. The idea was to collect first-hand information from decision makers, to discover if data play any role in their daily business operations and decision-making process. I was interested in knowing if there are any policies within the organizations for data collection, data storage and if the organizations valued the gathered data, thus utilising it for business enhancement.

The choice of organizations was random but the number of employees within organization was considered. No organization with less than 10 employees were contacted and only organizations with professional Chief Executive officers were considered for the research.

The survey was sent to organizations operating in sectors such as manufacturing, metal production, public service, IT service, utilities (water, gas, electricity), transportation, construction, commerce, marketing, financial services, food industry, and engineering. The numbers of employees within organizations varied from ten (10) to over hundred (100), and up to thousand employees for some organizations. Regarding the market these organizations operate or sell their products or services in Finland, some export, and some do both.

Concerning the information about data collection, the respondents were asked if their organizations collect internal data, meaning data about employees, system logs, and operating machines reports. Also, I wanted to know if the respondents collect any data from outside sources, e.g., if they have systems in place to collect data from outside their organizations. Additionally, the questions about tools being used to collect both internal and external data were asked. Furthermore, the questionnaire survey contained questions about data storage while the CEOs were asked about how the collected data are being stored. Are the data collected processed and sorted before being stored or are they stored in one repository before being processed and sorted for different usage?

Moreover, the survey contained questions about data dependency and future commitment to data collection. The idea was to determine if the decision makers actually utilize the data collected and use the insight gained in their daily decision-making processes. I also wanted to know if organizations have made data collection a priority and if they are committed to continuing collecting data for better business understanding and informed decision making.

4.3 Survey analysis

The research report is completely based on the strict unstructured questionnaire survey conducted and is specifically gotten from the thoughts and expressions of the respondents. The respondents could freely express themselves using the six themes mentioned in the research questions as guidelines. There were also possibilities for respondents to freely express their views on the topic in the last question. The Chief Executive Officers in various organizations provided me with key information about the impact and importance of gathering data in order to make informed decisions, thus gaining, and maintaining competitive advantages.

The respondents hold the position of CEOs and were chosen from different organizations operating in different sectors. The confidentiality of the respondents was assured in the research publication.

The questionnaire survey was sent by e-mail to the 63 CEOs representing different organizations operating in various sectors as mentioned earlier. Prior to sending the e-mail, I called the CEOs to inform them and explained what the research was all about.

Out of the 63 organizations, 44 responded. Among the 44 respondents, 27 organizations (61 percent) operate in manufacturing sector, 2 organizations (4.5 percent) operate as IT service providers, 2 organizations (4.5 percent) operate as public service providers, 2 organizations (4.5 percent) operate in energy sector, 2 organizations (4.5 percent) operate in transportation sector, while the rest 9 different organizations (20 percent) operate in construction business, wholesaling, marketing, accounting and payroll, restaurant, finances, engineering, and production of entrepreneurs.

As table 3 shows, 8 organizations had 10-20 employees, 3 organizations had 20-30 employees, 7 organizations had 30-40 employees, 3 organizations had 40-50 employees, 7 organizations had 50-100

employees, while 15 organizations had over 100 employees on their payroll lists. Noticeably, one respondent did not specify the number of employees.

TABLE 3. Distribution of the number of employees within the respondent organizations (n=43)

	n	Percent
10 to 20	8	18.6%
20 to 30	3	6.9%
30 to 40	7	16.3%
40 to 50	3	7.0%
50 to 100	7	16.3%
Over 100	15	34.9%

Among the respondents, 20 organizations have Finland as their main market, 12 organizations have export as their main market, and 12 have both Finland and export as their main markets.

Considering the data collection, which is the main theme of this thesis, an overwhelming majority (88 percent) of respondents collect various internal data, while 12 percent responded that they do not collect any internal data. Additionally, 73 percent of respondents admitted that they collect internal data from various operational equipment's system logs, while 27 percent do not collect data from systems logs. The organizations were asked to choose and name some tools and sources they use to collect internal data and table 4 below illustrates this. The most frequently used tools are ERP systems and tools used for financial reporting followed by sales reports, human resource management systems and interpersonal collaboration with employees.

TABLE 4. Tools for internal data collection

	n	Percent
Human Resources Department	22	50.0%
Information systems (ERP)	36	81.8%
Personal interaction with colleagues	24	54.5%
Company policy	8	18.2%
Financial reports	31	70.5%
Sales reports	26	59.1%

Pertaining to sources used to generate internal data, respondents were asked to select and name some of the sources used for internal data collection. 30 organizations selected employees' feedback option, while 29 organizations selected manufacturing data as one of the sources. Further 21 organizations selected various equipment devices logs as one of the sources. The option for production machines sensors was selected by 18 organizations, IOT devices (smartphones, and handheld devices) were selected by 15 organizations. Additionally, specific sources such as commercial /ERP systems, activity reporting into ERP, CRM, various statistical reports, surveys from employees, customer SaaS usage, and Idun application were also mentioned.

Regarding external data collection, out of the 44 respondents, 37 admitted that they collect external data compared to 7 who responded that they do not. As indicated in table 5 below, the main sources used are customers' feedback, suppliers' and partners' feedback, market trends, market research, website data collectors, and various external collaborations sources.

TABLE 5. Sources for external data collection

	n	Percent
Market Research	24	55.8%
Market Trends	28	65.1%
Website data collectors	22	51.2%
Customers' feedbacks	39	90.7%
Suppliers' and partners' feedbacks	32	74.4%
External collaboration	17	39.5%

Pertaining to data storage habits or practices in organizations, the respondents were asked to choose between given options for the data storage in their respective organizations. 22 organizations (51 percent) stated that they store both unstructured and structured data in one repository and extract needed data from there. 9 organizations (21 percent) replied that the data are collected and stored in a common repository before being extracted for analysis. Further 7 organizations (16 percent) admitted that they store data in separate repositories and extract needed data for analysis from there. Finally, 5 organizations (11 percent) stated that the data are collected and structured before it is stored.

Data collection and storage without any specific purpose is both a waste of resources and time for organizations. Therefore, the questions about insights and data dependency were posed to determine if there are benefits in gathering data, and if the organizations utilize the insights gained quotidianly in

their decision-making processes. 30 organizations (68 percent) admitted that the insights gained from collected data assist their organizations in decision making processes compared to 14 organizations (32 percent) who replied that the influence of gained insights depends on the matter at hand. However, it is necessary to note that not a single organization dismissed the significance of insights gained from collected data.

Regarding dependency on data collection, storage and analysis, the organizations were asked how dependent their organizations are on collected data. 27 organizations (61 percent) answered that they were dependent. 16 replied (36 percent) that their organizations were very dependent. However, 1 organization answered that it is not at all dependent.

Additionally, organizations were asked about the advantages in making informed decisions based on gained insights from collected data. As shown in table 6, it is clear that the majority of organizations acknowledge the advantages in making decisions which are informed and not based on feelings or intuitions.

TABLE 6. Advantages in informed decision making based on collected data (n = 44)

	n	Percent
Yes	38	86.4%
Not	0	0.0%
It depends	6	13.6%

The respondents were finally asked about their commitment to data collection in order to gain and maintain competitive advantage in the future. As table 7 shows, the overwhelming majority of organizations (93 percent) were certain that they are committed to continue gathering data in the future. Only one organization was categorically not committed to data collection in the future.

TABLE 7. Commitment to data collection in the future

	n	Percent
Yes	40	93.0%
No	1	2.3%
Not really sure	2	4.7%

The above research results demonstrate that data collection is not only limited to global mega organizations, but organizations in the regions of Pietarsaari and Kokkola also collect various types of data daily. Data collection is an ongoing mega trend that is here to stay. Irrespective of organizations' size or sector of operation, data play a major role, and it can be said that almost all organizations have data. Organizations in the regions of Pietarsaari and Kokkola have understood the importance of data in order to gain insights from it and in return make informed decisions. The decisions made from gained insights always have a good ground because the analysed data always illuminates the actual positions, conditions, and needs of an organization. The organizations know that by collecting data from different sources within and outside help them to improve services and products.

Organizations understand, that being aware of and knowing what the employees desire and value most will motivate them in their job performance and their overall wellbeing. Knowing what kind of trainings employees need will prove to be a perfect way of investing for the future and it gives a good picture of what skills are needed for future recruitments. The employees are the organizations' eyes and ears because they know what is happening on the operation floor and can give valuable and irreplaceable information to the decision makers. Employees are the ones operating the machines, they are the ones in contact with different departments within the organization. Furthermore, since the employees have contact with suppliers, customers, and competitors, gaining insights from them will benefit organizations in a positive manner.

Gathering data from operating machines is a good source of insights for organizations and this is not an exception for the respondents. The respondents demonstrate that logs and all other data from machines are important, and they collect data to learn about their operations and functionalities. Collecting data from machines allows decision makers to optimize their operations thus plan according to available machine capacity. In this way organizations can make timely decisions on investments and plan manpower and adjust their order stock.

Moreover, the organizations in the Pietarsaari and Kokkola regions seem to be up to date in their active actions on data collection. Different organizations have understood that there are many other technological tools such as website cookies, ERP information systems, CRM to facilitate data collection for organizations. Collected data can be very important because if well optimized can provide good insights. Tools are there to assist decision makers, however, feedback from customers and suppliers can prove irreplaceable for the success of businesses.

The respondents recognize that customers' behaviour has changed in the past decades. They understand that the time of developing services and products in hope that customers will buy them is almost over. Recently the trend has been that customers and sellers collaborate for new products or service design and development. Organizations perform better if they understand their customers' behaviour and if they can forecast and be ahead of trends.

4.4 Findings

The overwhelming majority of organizations in the researched regions clearly indicated that they collect both internal and external data. They understood what is meant by internal data because having insights of what is happening internally facilitates overall decision-making process within organizations. Information from human resource department, feedback from employees, and information collected from internal machines and sensors prove to be vital for organizations.

The other revelation was that organizations are not only relying on internal data, but external data is also important for organizations as well. There are measures in place to gather data from external source as to have insights from customers, distributors, competitors, market trends and demands. Various technological tools are used to collect the vital and important data for better decision-making processes. The respondents seek ways to collect data and are using data to be informed on various situations and circumstances.

It also became clear that organizations in the studied regions are technological minded and thrive forward by investing both money and resources on technologies that enable them to collect and store data safely. From the given answers, it can be said that some used the concept of a data lake, which is a concept of storing both structured and unstructured data before it is being extracted for analysis. The other organizations separate data before it is being stored. The results also indicated that technologies such as data warehouses, data marts, or data cubes are in use at some organizations.

Further analysis of the results reveals that all the organizations that participated on the survey are collectors of internal data, however, when it came to external data collection, 13 percent of manufacturing organizations answered that they do not collect any external data. It can be assumed that these organi-

zations do not have tools for data collection, or they do not have policies or system in place for external data collection. It can also be assumed that these companies do not understand the value gained from external data. More importantly, the research confirms the impact information systems and digitalization have on organizations.

Many tools and means are used for internal and external data collection. Various ERP systems provide immense information for managers and informative reports can be customized and produced for specific needs. ERP systems have become an integral tool for most of organizations even among the participants, ERP was named as a great source of data. Many organizations rely more on their financial reports to determine their performance, but we all know that financial reports reflect always to past performances and not the future. However, the insight gained from financial performance can be helpful for future planning and trends adaptation.

In another development, the utility (gas, water, and electricity) sector seems to be a forerunner in regard to the usage of IoT technology. All the respondents in the utility sector utilise the technology more than their counterparts. The reason for this could be financial muscles and technological demands in order to be updated and have information on real time basis. Employees' feedback is good source for information, and this is understood by manufacturing and public service organizations. However, IT service providers seem to be equally divided on this source of data, but the utility sector are categorically not interested in collecting data from their employees. This is very interesting finding. What are the reasons behind this? Are feedback from employees not relevant for this particular sector? Are other tools more reliable than employees' feedback?

An interesting finding is that utility sector seems to be technologically minded when it comes to data collection. This sector collects huge amount of data from machines and equipment in comparison to their counterparts. Moreover, manufacturing sector seems to be the more balanced of them all because they combine both technological means and human inputs in their business operations.

A pleasing development is that customers' feedback is very appreciated and are utilised in all business sectors. The saying that customer is king is demonstrated by all the respondents. Organizations listen to their customers for improvement suggestions, compliments and they seem to prioritize the customers. The respondents have understood the role the customers play in any business today and are com-

mitted to listen to them and to meet their demands. Market research is a new way of studying the customers' behaviour and a means by which organizations can design their services and products according to the market. It seems like the respondents are still finding their way in market research mindset. However, manufacturing sector seems to have come further than its counterparts. Why is this? Is it because the majority of them export their products? Or is it because competition is hard thus, innovation is inevitable?

Means, such as market trends, are in use and external collaborations are finding their way in many organizations, compared to website data collectors which is used by IT service providers and the public service sector. Moreover, all the respondents are really good at collaborating with their suppliers and partners. Collaborating with partners and suppliers assists in products and services design because these provide vital information and help in delivering good customer service.

Organizations in manufacturing, public service and IT service providers are committed to data collection in the future to gain and maintain competitive advantages. This is very promising because data collection and processing is not only today's practice. Data have positioned themselves as the most vital asset many organizations possess. The more data organizations own about their employees, customers, suppliers, partners, and competitors the better they are equipped for current demands and future challenges and opportunities.

However, respondents in utility sector seem to be satisfied with what they have, therefore they answered that they are not at all committed to data collection in the future. What lays at the heart of their uncertainty? Could it be that the market is already regulated on this sector that it does not need any more data? Or could it be that the respondents did not want to answer favourably because of the GDPR (general data protection regulation)? One thing is certain for all organizations, data are an integral part of their daily business. Lack of data can be compared to surfing in the dark, and this can be detrimental to the organizations.

Knowledge derives from information, information in turn derives from processed data. Data by definition is a collection of numbers, letters, pictures, audios, videos, and other symbols. For organizations, data are very important, this was repeatedly emphasised in discussions with most of respondents. They all agreed that they collect various types of data, and they acknowledged that unprocessed data are almost meaningless.

Some of the challenges are knowing what kinds of data are important and where they are stored in order to make use of them when needed. An example was given about tools to extract data with and data visualization skills and then analysis was the other named challenge. How can organizations collect and organize data for further analysis in other to gain insight is a common challenge? Another challenge for the organizations was to produce holistic reports. Organizations dealing with data today produce separate reports for each business module in the systems. Possessing tools that facilitate a holistic reporting are still missing in many organizations I discussed with.

Finally, the research confirmed that organizations in this study are dependent on insights gained from data and are committed to data collection in the future. The following are the words from one of the respondents who stated that "I personally believe that all companies in the future will depend on data driven decision making as the market develops and changes to become completely service driven." Collected data enable organizations to have vital and relevant information, thus supporting decision makers to make informed decisions.

4.5 Reliability and validity

This thesis work can be trusted because of the first-hand information and responses received from the respondents who are actual representatives of their organizations. It is right for me to reiterate that the research is a qualitative study on recognizing the value of data in business operations, a study on gathering internal and external data and ways to utilize it in business strategies.

The respondents are all chief executive officers / Managing directors of both profit, and none profit organizations in the regions of Pietarsaari and Kokkola. These organizations operate in different fields which makes them diverse in knowledge, interest, and perspective about the various themes of discussion. The number of respondents and their different fields of operations coupled with various sizes of the organizations is something that can add up to the validity of this research.

It can also be said about the thesis that the necessary methodologies of a research process were followed. Processes such as the identification of the research problem, the revision of the literature, the specification of the research purpose, the collection of data, the process of data analysation and interpretation, and finally the reporting and evaluating the research.

5 CONCLUSION AND DISCUSSION

As stated in the introduction, the thesis process began by becoming familiar with the topic of data collection, data storage, and utilizations of gained insights for better business decision making process. The research was to determine if organizations in Pietarsaari and Kokkola regions collect data for analysis in order to gain and maintain competitive advantages. I wanted to determine if organizations actually have policies in place for data collection and if they see advantages in data. The scope of the thesis was limited to the collection, storage and frequent usage of data and commitment to future data collection and data utilisation. The storage repository method called the data lake was also considered in detail for this thesis.

Data are what most companies desire to have. Knowing the current business situation and being able to forecast the future trends and adapt the business models accordingly is what organizations thrive for. To be able to manufacture products or design services based on actual demands from customers will mean that organizations will not assume but they will operate in reality. The terms data driven, and data infused businesses has become so common that their meanings are even lost. Organizations want data and should make efforts to have as much data as possible and to extract helpful information for better decisions and brighter future.

Organizations need to collect right kind of data in order to get quality insights that can benefit them in the decision-making processes. Bad quality data will hurt organizations in the long run instead of doing any good. Decisions based on bad sources are as dangerous for organizations as having no data at all. As this questionnaire correctly points out, data can be anything from employee feedback to purchased data collection from a third party to machine performance data. All data can be useful. There will always be a need to find a balance. Too little data can lead to resource wastage, too much can lead to "paralysis by analysis". In all data analysis, one needs to be aware that data usually refers to past performance and is usually analysed with past experience according to past constraints. We should remember that much of the future we are still free to define.

Right tools for data collection are vital in the data collection process. Organizations must pay careful attention to tools used for data collection. Bad tools mean bad quality data and probably old and unnecessary data for organizations. There are many software solutions that assist organizations to collect timely and good quality data. Any negligence in investing time and money in these good tools can be

dangerous for organization. Inaccurate tools will lead decision makers astray thus resulting in inaccurate business processes. Moreover, organizations must also pay attention to safe and reliable storage solutions to guaranty the reliability of the data they are dealing with. Contaminated or unreliable storage solutions are dangerous for organizations.

As collecting and storing data are important to organizations so is the process of extracting good data for analysis. After data are collected and stored safely, the process of extracting the needed data begins. The challenge as of today, as one of the respondents in this survey claimed, is to combine different sources of data and make a dynamic reporting. Organizations need clear policies and reliable tools to extract good quality data for further usage. There are smart applications that assist in data extraction, and these should be used for better quality and faster processing.

Besides the above-mentioned aspects, skills for analysing, reporting and visualization are needed. Knowledgeable data analysts are required to bring to light the value of data itself. There is no need for investing in great tools for data collection, storage and extraction while lacking the needed skills for interpretation and visualization. The started process must be completed for organizations to fully benefits from data. Appropriate trainings for the staff dealing with data is always a good investment because it always pays off when companies learn to use data to learn the past, know the present and forecast the future.

There are benefits connected to the usage of data in organizations. Organizations use the gained knowledge to learn from the past and plan for the future. Insights gained provide various lessons for decision makers, for example, if a department planned poorly in the past, the gained data could show what went wrong and why. Data if well analysed and interpreted will provide the future trends, it can point the direction the business should take based on customers' behaviour and demands. Data bring benefits such as exposing the lack of expertise within the organization and helps to make right decisions.

To survive today, organizations must take data and their benefits seriously. As one of the respondents pointed out, possessing large amounts of data is not sufficient. The usage of collected data (from one or different sources) must be improved to fully get a complete picture of the organization. The respondent explained that different departments within the organization collect data and analyse data to gain insights related to their specific departments. The respondent proceeded by saying that to fully benefit from data, the next step for the organization is to form a team of skilful data analysts with both

technological and commercial expertise. These analysts' job would be to analyse all the organization's data and produce a combined report from all data sources and departments. This will provide a better picture because all the insights will be presented as a common report.

The organizations in Pietarsaari and Kokkola regions show that they are doing all they can to stay competitive through knowledge gained from data. They are also committed to gather, store, and analyse data to enhance their expertise and development, thus, gaining and maintaining competitive advantages. It can be said that having data is holding the keys to future success.

For the implication and managerial recommendation, the concept of information systems (IS) which consists of people, technology, processes, and data should be at the heart of every organization decision making process. For many organizations investing in tools to accomplish a task proceeds the most important step of first defining the process. Processes need to be thought through and defined before the implementation of any tools. This is not an exception with data, processes for good data management and handling need to be present before data collection, storage, visualization, and analysis.

Data can become a burden for many organizations if good care is not considered. The consequences of poor data management can be likened to someone possessing gold but begging for bread instead. Data should become everybody's business within organizations. Organizations should cultivate the Lean organization model mindset that includes data collection and data handling. Every employee should be aware of value that comes from data, and they should be good stewards of data. The mistake of thinking that duties related to data are only for IT professionals and Finance department should be completely eliminated. If your organization possess gold then, implement policies for good stewardship.

The other way to make data useful is by implementing good knowledge management system and procedure and making sure that every employee is trained for proper handling of data and information extracted thereafter. Also, good business intelligence technic (BIT) such as data mining, deep learning, and machine learning are equally important as the collection of data. Training in BIT and data analysis should not be neglected. Being able to collect quality data, storing it safely, sorting it before extracting it for visualization and analysis should be the goal for any modern organization. Make data handling your organizations' priority. Remember that timeliness, accuracy, and completeness are the characteristics that make information valuable.

This research focussed on the collection of both internal and external data and the storage thereafter. However, many more topics can be researched for the future. Topics such as skills needed for good data storage and processes for cleaning, extracting, visualizing, and analysing data. The usage of technologies for business intelligence such as data mining, deep learning, and machine learning could be a good topic for research.

REFERENCES

Alghushairy, O. & Ma, X. 2019. *Data storage*. Available at: https://www.researchgate.net/publication/335754159_Data_Storage. Accessed 21.6.2021.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B. & Kochut, K. 2017. *A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques*. Available at: https://www.researchgate.net/publication/318336890_A_Brief_Survey_of_Text_Mining_Classification_Clustering_and_Extraction_Techniques. Accessed 01.7.2021.

Altheide, C. & Carvey, H. 2011. *In Digital Forensics with open-source tools: Metadata extraction*. Available at: https://www.sciencedirect.com/topics/computer-science/metadata. Accessed 1.7.2021.

Arthur, L. 2013. *Big data marketing: Engage your customers more effectively and drive value*. Available at: ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/cop-ebooks/detail.action?docID=1426518. Accessed 3.7.2021.

Australian Bureau of Statistics 2020. *Statistical language: Quantitative and Qualitative data*. Available at: https://www.abs.gov.au/websitedbs/D3310114.nsf/Home/Statistical+Language+-+quantitative+and+qualitative+data. Accessed 29.6.2021.

Baud, N., Franchot, A. & Roncalli, T. 2002. *Internal Data, External Data and Consortium Data: How to mix them for measuring operational risk*. Available at: https://www.researchgate.net/publication/251242277_Internal_Data_External_Data_and_Consortium_Data_-_How_to_Mix_Them_for_Measuring_Operational_Risk. Accessed 5.7.2021.

Bespalov, A., Michel, M. C. & Steckler, T. 2020. *Good research practice in non-clinical pharmacology and biomedicine. Handbook of Experimental pharmacology, volume 257*. Available at: https://link.springer.com/chapter/10.1007/164_2019_288. And https://link.springer.com/book/10.1007/978-3-030-33656-1. Accessed 8.7.2021.

Blumzon, C. F. I. & Pănescu, A. T. 2019. *Good research practice in non-clinical pharmacology and biomedicine*. Available at: https://www.researchgate.net/publication/337691364_Data_Storage. Accessed 30.5.2021.

Cambridge Assessment International Education 2017. *Data, information and knowledge*. Available at: https://www.cambridgeinternational.org/Images/285017-data-information-and-knowledge.pdf. Accessed 03.6.2021.

Checkland, P. & Holwell, S. 1998. *Information, systems and information systems: Making sense of the field*. Available at: C:\PhD\Chapters\AlexaChpt3.wpd (up.ac.za). Accessed 15.5.2021.

Daniel, L. E. & Daniel, L. E. 2012. *Digital forensics for legal Professionals*. Available at: https://www.sciencedirect.com/topics/computer-science/metadata. Accessed 25.6.2021.

Eiras, J. R. 2011. *Data collection and storage.* Available at: https://ebookcentral-proquest-com.ezproxy.centria.fi/lib/cop-ebooks/reader.action?docID=3021663&query=EIRAS. Accessed 22.6.2021.

Faniel, I. M., Kriesberg, A. & Yakel, E. 2015. *Social Scientists' satisfaction with data reuse*. Available at: https://www.oclc.org/content/dam/research/publications/2015/faniel-kriesberg-yakel-2015-social-scientists-satisfaction-preprint.pdf. Accessed 3.6.2021.

Gartner, T. 2008. *Kernels for structured data, World Scientific Publishing Company, 2008*. Available at: ProQuest Ebook Central, http://ebookcentral.proquest.com/lib/cop-ebooks/detail.action?docID=1193194. Accessed 3.6.2021.

Grossman, R. L. 2019. *Data lakes, clouds, and commons: a review of platforms for analysing and sharing Genomic data.* Available at: https://www.sciencedirect.com/science/article/abs/pii/S0168952518302257. Accessed 28.6.2021.

Han, J. & Pei, J. 2012. *Data mining (Third edition): Data warehousing and online analytical processing*. Available at: https://www.sciencedirect.com/topics/computer-science/data-cube. Accessed 05.7.2021.

Han, J. & Kamber, M. 2006. *Data mining: Concepts and techniques (second edition)*. Available at: https://books.google.fi/books?id=AfL0t-YzOrEC&pg=PA189&dq=Data+cube&hl=en&sa=X&ved=2ahUKEwic4sX17sPxAhWDHXcKHemPAcIQ6AEwAHoECAYQAg#v=onepage&q=data%20cubes&f=false. Accessed 20.7.2021.

Hay, D. C. 2006. *About metadata models*. Available at: https://www.sciencedirect.com/topics/computer-science/metadata-repository. Accessed 15.6.2021.

Hobbs, L. & Smith, P. 2005. *Oracle 10 g data warehousing: building a data warehouse poses many challenges*. Available at: https://www.sciencedirect.com/topics/computer-science/data-warehouses. Accessed 10.6.2021.

Inmon, W. H. & Linstedt, D. 2015. *Data architecture: A primer for data scientist*. Available at: https://www.sciencedirect.com/topics/computer-science/data-warehouses. Accessed 21.6.2021.

Johnston, L. R. 2017. *Curating research data volume one: practical strategies for your digital repository*. Available at: https://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988596_crd_v1_OA.pdf. Accessed 08.6.2021.

Kabir, S. M. S. 2016a. *Basic guidelines for research: An introductory approach for all disciplines*. Available at: https://www.researchgate.net/publication/325846733_INTRODUCTION_TO_RESEARCH. Accessed 2.6.2021.

Kabir, S. M. S. 2016b *Methods of data collection*. Available at: https://www.researchgate.net/publication/325846997_METHODS_OF_DATA_COLLECTION. Accessed 6.6.2021.

Kleppmann, M. 2017. *Designing Data-intensive Applications*. Available at: https://www.oreilly.com/library/view/designing-data-intensive-applications/9781491903063/ch01.html. Accessed 16.6.2021.

Liu, K. & Dong, L. 2012. *Research on cloud data storage and its architecture implementation*. Available at: https://www.researchgate.net/publication/257723864_Research_on_Cloud_Data_Storage_Technology_and_Its_Architecture_Implementation. Accessed 19.6.2021.

Llave, M. R. 2018. *Data lakes in business intelligence: Reporting from the trenches*. Available at: https://www.sciencedirect.com/science/article/pii/S1877050918317046. Accessed 26.6.2021.

Loshin, D. 2013. *Business Intelligence (second edition): Data warehouses and the technical business intelligence architecture*. Available at: https://www.sciencedirect.com/topics/computer-science/data-repository. Accessed 01.6.2021.

Marinescu, D. C. 2018. *Cloud computing second edition: Big data, data streaming, and the mobile cloud*. Available at: https://www.sciencedirect.com/topics/computer-science/data-repository. Accessed 15.5.2021.

McKinsey Global Institute 2016. *The age of analytics: Competing in a data-driven world*. Available at: https://www.mckinsey.com/~/media/mckinsey/industries/public%20and%20social%20sector/our%20insights/the%20age%20of%20analytics%20competing%20in%20a%20data%20driven%20world/mgi-the-age-of-analytics-full-report.pdf. Accessed 7.6.2021.

Microsoft Corporation 2021. *Online transaction processing (OLTP).* Available at: https://docs.microsoft.com/en-us/azure/architecture/data-guide/relational-data/online-transaction-processing. Accessed 12.5.2021.

Miloslavskaya, N. & Tolstoy, A. 2016. *Big Data, Fast Data and Data Lake Concepts*. Available at: https://www.sciencedirect.com/science/article/pii/S1877050916316957. Accessed 29.6.2021.

Plotkin, D. 2021. *Data stewardship (second edition): Practical data stewardship, a metadata repository*. Available at: https://www-sciencedirect-com.ezproxy.centria.fi/science/article/pii/B9780128221327000103. Accessed 30.6.2021.

Praveen, S. & Chandra, U. 2017. *Influence of structured, semi-structured, unstructured data on various data models*. Available at: https://www.researchgate.net/profile/Umesh-Chandra-8/publication/344363081_Influence_of_Structured_Semi-_Structured_Unstructured_data_on_various_data_models/links/5f6c6ee7a6fdcc0086386767/Influence-of-Structured-Semi-Structured-Unstructured-data-on-various-data-models.pdf. Accessed 28.5.2021.

Qamar, S. 2015. *Data mapping for data warehouse design*. Elsevier Science & Technology. ProQuest Ebook Central. Available at: http://ebookcentral.proquest.com/lib/cop-ebooks/detail.action?docID=4202933. Accessed 14.6.2021.

Ratcliffe, S. 2018. *Oxford Essential Quotations (6 ed).* Available at: https://www.oxfordreference.com/view/10.1093/acref/9780191866692.001.0001/q-oro-ed6-00019739. Accessed 7.6.2021.

Reeve, A. 2013a. *Managing data in motion: Data warehouse and operational data store*. Available at: https://www.sciencedirect.com/topics/computer-science/data-warehouses. Accessed 07.6.2021.

Reeve, A. 2013b. *Managing data in motion: Batch data integration architecture and metadata*. Available at: https://www.sciencedirect.com/topics/computer-science/data-warehouses. Accessed 20.6.2021.

Schatsky, D., Camhi, J. & Muraskin, C. 2019. How third-party information can enhance data analytics. Available at: https://www2.deloitte.com/us/en/insights/focus/signals-for-strategists/smart-analytics-with-external-data.html. Accessed 16.6.2021.

Singman, P. 2021. *Data lakes: The definitive guide*. Available at: https://lakefs.io/data-lakes/. Accessed 15.6.2021.

Sivarajah, S., Kamal, M. M., Irani, Z. & Weerakkody, V. 2017. *Critical analysis of big data challenges and analytical methods*. Available at: https://www.sciencedirect.com/science/article/pii/S014829631630488X. Accessed 15.7.2021.

Tandy, J., Ceolin, D. & Stephan, E. 2016. *CSV on the web: Use cases and requirements.* Available at: https://www.w3.org/TR/csvw-ucr/. Accessed 16.7.2021.

Techopedia 2012. *Data storage*. Available at: https://www.techopedia.com/definition/23342/data-storage. Accessed 16.5.2021

Tupper, C. D. 2011. *Data architecture: Dimensional databases from enterprise data models*. Available at: https://www.sciencedirect.com/topics/computer-science/data-warehouses. Accessed 19.6.2021.

Wallace, P. 2018. *Introduction to information systems third edition*. New Jersey: Pearson Education.

Vaughan, J. 2019. *Definition of data*. Available at: https://searchdatamanagement.techtarget.com/definition/data. Accessed 19.7.2021.

Wellington, J. & Szczerbinski M. 2007. *Research Methods for the Social Sciences*. London: Bloomsbury Publishing Plc.

Whitney, C.W., Lind, B.K. & Wahl, P.W. 1998. *Quality assurance and quality control in longitudinal studies. Epidemiologic Reviews.* Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.618.53&rep=rep1&type=pdf. Accessed 19.7.2021.

Wilson, E. 2019. *Forecaster's & Planner's Guide to Data*. Available at: https://demand-planning.com/2019/08/26/forecasting-data-types/. Accessed 19.7.2021.

Xie, Q., Wang, J., Kim, G., Lee, S. & Song, M. 2021. *A sensitivity analysis of factors influential to the popularity of shared data in data repositories*. Available at: https://www-sciencedirect-com.ezproxy.centria.fi/science/article/pii/S1751157721000134. Accessed 25.6.2021.

Ziegler, J., Reimann, P., Keller, F. & Mitschang, B. 2020. *A graph-based approach to manage CAE data in data lake*. Available at: https://www-sciencedirect-com.ezproxy.centria.fi/science/article/pii/S2212827120310349. Accessed 21.6.2021.

Zins, C. 2005. *What is the meaning of data, information and knowledge?* Available at: https://pciucr.files.wordpress.com/2011/03/what-is-the-meaning-of-data.pdf. Accessed 30.5.2021.

Recognizing the value of data in business operations

About the company

1. In which business sector does your company operate?

- ○ Manufacturing
- ○ Metal production
- ○ Public service
- ○ IT service
- ○ Utilities (water, gas, electricity)
- ○ Transportation
- ○ Construction
- ○ Other, please specify

2. Number of employees

- ○ 10 to 20
- ○ 20 to 30
- ○ 30 to 40
- ○ 40 to 50
- ○ 50 to 100
- ○ Over 100

3. What is your main market

- ◯ Finland
- ◯ Export
- ◯ Both Finland and Export

Questions about Internal Data. Knowing what is happening inside your company gives you insights and enables you to plan well and make informed decisions

4. Does your company collect information about the employees (career paths, expertise, skills, and needs for further training)?

- ◯ Yes
- ◯ No

5. Does Your company collect information from equipments' system logs?

- ◯ Yes
- ◯ No

Sources used to collect internal data

6. What are the tools, sources used to collect information? (Please choose one or more answers)

- ☐ Human Resources Department
- ☐ Information systems (ERP)
- ☐ Personal interaction with colleagues
- ☐ Company policy
- ☐ Financial reports
- ☐ Sales reports

7. From which of the following does your company collect data for day to day business operations? (Please choose one or more answers)

- [ ] Production machines sensors
- [ ] IoT devices (smartphones, handheld devices,...)
- [ ] Employees' feedback
- [ ] Manufacturing data
- [ ] Equipment device logs
- [ ] Others, please specify

Questions about External Data. Knowing what is happening outside your company gives you insights and enables you to plan well and make informed decisions

8. Does your company collect data from external sources to make business decisions?

- ◯ Yes
- ◯ No

9. Which of the following external sources does your company use to collect data?

- [ ] Market Research
- [ ] Market Trends
- [ ] Website data collectors
- [ ] Customers' feedbacks
- [ ] Suppliers' and partners' feedbacks
- [ ] External collaboration

Collected Data storage before processing and analysis

10. Which of the following describes the data storage in your company?

- ◯ The data is collected and structured before being stored
- ◯ The data is collected and stored in a common repository before being extracted for analysis
- ◯ We store both structured and unstructured data in one repository and extract needed data from there
- ◯ We store data in separate repository and extract needed data for analysis from there

11. Do the insight gained from collected data assist your company in decision making process?

- ◯ Yes
- ◯ No
- ◯ Both Yes and No

12. How dependent is your business on the collected data?

- ☐ Very dependent
- ☐ Dependent
- ☐ Not at all dependent

13. Do you see advantages in making informed decisions based on gained insights?

- ◯ Yes
- ◯ Not
- ◯ It depends

14. Is your company committed to continue collecting data to gain and maintain competitive advantage?

- ◯ Yes
- ◯ No
- ◯ Not really sure

15. Do you have anything to add or comment on this topic?

APPENDIX 2

The statistical report of the questionnaire survey

Statistics

Question	Count	Average	Confidence interval	Median	Standard deviation
1. In which business sector does your company operate?	44	3.11	2.24 – 3.98	1	2.94
2. Number of employees	43	4	3.42 – 4.58	5	1.94
3. What is your main market	44	1.82	1.57 – 2.07	2	0.84
4. Does your company collect information about the employees (career paths, expertise, skills, and needs for further training)?	43	1.12	1.02 – 1.21	1	0.32
5. Does Your company collect information from equipments' system logs?	44	1.27	1.14 – 1.41	1	0.45
6. What are the tools, sources used to collect information?(Please choose one or more answers)	147	3.46	3.18 – 3.75	3	1.77
7. From which of the following does your company collect data for day to day business operations? (Please choose one or more answers)	122	3.39	3.12 – 3.65	3	1.47
8. Does your company collect data from external sources to make business decisions?	44	1.16	1.05 – 1.27	1	0.37
9. Which of the following external sources does your company use to collect data?	162	3.48	3.24 – 3.73	4	1.59
10. Which of the following describes the data storage in your company?	43	2.72	2.46 – 2.98	3	0.88
11. Do the insight gained from collected data assist your company in decision making process?	44	1.64	1.36 – 1.91	1	0.94
12. How dependent is your business on the collected data?	44	1.66	1.5 – 1.81	2	0.53
13. Do you see advantages in making informed decisions based on gained insights?	44	1.27	1.07 – 1.48	1	0.69
14. Is your company committed to continue collecting data to gain and maintain competitive advantage?	43	1.12	0.98 – 1.25	1	0.45

APPENDIX 3

FREE FORMULATED COMMENTS FROM RESPONDENTS ON THE TOPIC

Do you have anything to add or comment on this topic?

Responses
As this questionnaire correctly points out, data can be anything from employee feedback to purchased data collection from a third party to machine performance data. All can be useful. There will always be a need to find a balance. Too little data can lead to resource wastage, too much can lead to "paralysis by analysis". In all data analysis, one needs to be aware that data usually refers to past performance and is usually analysed with past experience according to past constraints. We should remember that much of the future we are still free to define.
For the moment we have access to data from different sources. Typical report on sales and profits are ok and can be seen in reporting systems. The challenge as of today is to combine different sources of data (capacity, availability of material, forecast, …) and make a dynamic reporting (not to extract and combine in excel)
I personally believe that all companies in the future will depend on data driven decision making as the market develops and changes to become completely service driven.
Collecting data is a normal daily task in our business (utilities)