Enabling data-driven decision-making for a Finnish SME: a data lake solution

Master of Science in Technology Thesis
University of Turku
Department of Computing
Software Engineering
2023
Marika Helttula

UNIVERSITY OF TURKU
Department of Computing

Marika Helttula: Enabling data-driven decision-making for a Finnish SME: a data lake solution

Master of Science in Technology Thesis, 61 p.
Software Engineering
April 2023

In the era of big data, data-driven decision-making has become a key success factor for companies of all sizes. Technological development has made it possible to store, process and analyse vast amounts of data effectively. The availability of cloud computing services has lowered the costs of data analysis. Even small businesses have access to advanced technical solutions, such as data lakes and machine learning applications.

Data-driven decision-making requires integrating relevant data from various sources. Data has to be extracted from distributed internal and external systems and stored into a centralised system that enables processing and analysing it for meaningful insights. Data can be structured, semi-structured or unstructured. Data lakes have emerged as a solution for storing vast amounts of data, including a growing amount of unstructured data, in a cost-effective manner.

The rise of the SaaS model has led to companies abandoning on-premise software. This blurs the line between internal and external data as the company's own data is actually maintained by a third-party. Most enterprise software targeted for small businesses are provided through the SaaS model. Small businesses are facing the challenge of adopting data-driven decision-making, while having limited visibility to their own data.

In this thesis, we study how small businesses can take advantage of data-driven decision-making by leveraging cloud computing services. We found that the reporting features of SaaS based business applications used by our case company, a sales oriented SME, were insufficient for detailed analysis. Data-driven decision-making required aggregating data from multiple systems, causing excessive manual labour. A cloud based data lake solution was found to be a cost-effective solution for creating a centralised repository and automated data integration. It enabled management to visualise customer and sales data and to assess the effectiveness of marketing efforts. Better skills at data analysis among the managers of the case company would have been detrimental to obtaining the full benefits of the solution.

Keywords: Data-driven decision-making, SMEs, Cloud computing, Amazon Web Services, Data lakes

Contents

List of Figures

List of Tables

1 Introduction

In this thesis, we study data-driven decision-making in a small business context and provide a technical solution for overcoming the challenges related to data availability. In this introductory chapter, we first describe the phenomenon more generally, then from a small business perspective and finally conclude with the research design and questions.

1.1 Research phenomenon

Technological innovation of the past few decades has enabled business leaders to make critical decisions effectively based on solid proof instead of intuition. This is called data-driven decision-making. The amount of data easily available for managers has skyrocketed and will continue to grow at a fast rate. Insights derived from this data can help businesses become more successful. It can make businesses more profitable by increasing customer satisfaction, cutting supply chain costs and focusing efforts on activities with the biggest return. Furthermore, data-driven decision-making can help businesses grow and innovate by revealing new business opportunities.

There are many examples of companies that have made disruptive decisions based on data. Amazon went from being a bookstore to a multinational retailer of almost anything consumers might wish to buy. They invested heavily on IT infrastructure over the years, which made them effective, but also led them to pivot to a new

business sector, namely cloud computing services. They realised they could sell their excessive capacity to other companies who did not wish to maintain their own IT infrastructure. Today Amazon Web Services (AWS), established in 2006, is the most profitable business sector of the company [1].

Amazon was not only able to collect and utilise data to make critical business decisions, but in the process became a popular platform for other companies to take advantage of data-driven decision-making. Cloud computing services offer easy access to the storage and processing capacity needed to collect and analyse large amounts of data. The availability of low cost cloud computing services is especially important for small and medium sized companies wishing to make more sound business decisions without having to maintain a heavy IT infrastructure.

There are numerous solutions available for storing, processing and analysing data. Some of the traditional solutions for storing large amounts of data, like data warehouses, may not be suitable for big data and the growing amount of unstructured data. Data lakes are becoming increasingly popular as the technical solution for storing vast amounts of versatile data and the availability of cloud computing platforms have made them more accessible also for small businesses. Data lakes can be used to tackle big data challenges connected to the variety, volume and velocity of data, but also to reduce data silos that exist in all distributed systems. The rise of Software-as-a-Service (SaaS) as the distribution model for modern business software can hinder companies of all sizes from accessing their data flexibly. This increases the need for solutions such as data lakes for collecting data from various systems in order to ensure data availability.

1.2 Research context

Most small and medium sized enterprises (SMEs) operate in highly competitive markets. This is especially true for those companies that sell consumer goods and

services. In a competitive market, it is important to react quickly to changes in the business environment and make sound decisions. Successful companies are able to identify new opportunities and to take advantage of them faster than their competitors. The effective use of technology can be a success factor in this race. Small businesses, however, have more limited possibilities for investing in technology compared to their larger adversaries. They may lack both the financial resources and the technical capabilities needed for IT investments. Most SMEs are dependent on software and IT infrastructure maintained by third-party software providers.

Selecting the right business software provider can be a critical success factor for small businesses. SaaS applications typically cater for the needs of vast audiences. If the needs of the company align well with a large population of other actors then chances are that there is suitable software available for an affordable monthly subscription fee. The most commonly needed features are being developed by the service provider, so there is no need for expensive customisation as long as the company can settle for the existing features or patiently wait for new features to appear in time.

Data-driven decision-making requires access to data, be it internal like customer or sales data, or external like website traffic or social media entries. SMEs that use SaaS applications may not have flexible access to their own data. The data is typically stored on the service provider's servers. The users of the application have access to their data through the application as long as they remain subscribers, but may end up losing all their data once the subscription is terminated.

Most modern business applications offer an application programming interface (API) that can be used to integrate distributed systems. Automation of data transfer between distributed systems through APIs can reduce the need for costly manual data transfer steps. APIs can be utilised for enabling data-driven decision-making by transferring data from distributed systems to a single repository, a data lake for

instance, that is used for data analysis and visualisation.

The first step for data-driven decision-making is ensuring data availability. The next step is to make sure the right people have access to all the relevant datasets and can utilise them effectively. Large companies may have dedicated departments with skilled data analysts creating business intelligence (BI) reports for managers. Small business leaders are rarely data scientists nor can employ an army of specialists to do the data crunching. They need data discovery and analysis tools that enable getting started easily and becoming proficient in time through learning by doing. Data-driven decision-making needs to be rooted into the company culture in order to make it an integral part of the operational model of the company.

1.3 Research design and questions

This study strives to answer, how to enable data-driven decision-making in a Finnish SME by leveraging cloud computing, more specifically a cloud based data lake? This main research question can be broken into three sub-questions:

1. What are some of the technical challenges hindering data-driven decision-making in small businesses?

2. How can a cloud based data lake be utilised to solve these challenges, especially related to data integration, data visualisation and data analysis?

3. What kind of an effect does the implementation of a cloud based data lake have on the ability of the case company to take advantage of data-driven decision-making?

We approach the research questions from a design science methodology point of view. Following the fundamental principle of design science proposed by Hevner et al. [2], we aim at creating knowledge and understanding of the design problem and its

solution through building and application of an artefact, namely a cloud based data lake. We build a data lake solution using serverless components provided by Amazon Web Services. The focus of the study is on solving the challenges related to data availability. We acknowledge that a technical solution cannot be the sole enabler of data-driven decisions-making in any organisation. Most companies, especially smaller ones, need to work on their data analysis capabilities and organisational culture in order to fully leverage the advantages of data-driven decision-making. Yet, solving the issue of data availability is one of the first steps in the process of becoming a data-driven organisation and hence an important one.

The remainder of this thesis is divided into five chapters. In Chapter 2, we start by discussing, what data-driven decision-making is and what kind of benefits and challenges it might pose to small businesses? We also look at how technological development has changed the characteristics and amounts of data used in decision making and how this is shaping the requirements for data storage systems. We review some main differences between the more traditional data warehouse solution and the relatively new data lake concept, which is then discussed in more detail. Chapter 2 concludes with a discussion on the role of cloud computing for small businesses in general and for building data lakes in particular. Chapter 3 is dedicated to introducing the case company. We discuss the limitations of their current technological business environment in terms of data-driven decision-making and the objectives they have for taking advantage of data and the insights it might bring. Chapter 4 discusses the research methodology and gives a detailed view of the iterative research process. Chapter 5 describes the technological solution that was created to solve the case company's problems related to data availability and to reach their objectives in terms of data-driven decision-making. Finally, Chapter 6 concludes the findings of the study.

2 Data and decision support systems

2.1 Data-driven decision-making in SMEs

Data-driven decision-making refers to making decisions based on solid proof instead of assumptions. Technological innovations of the past decades have enabled companies to store and analyse vast amounts of data in support of their decision making processes.[3] Such data-driven companies are likely to be more productive than their peers. Studies have shown that the use of big data technologies correlates with significant additional productivity growth. Data-driven decision-making was associated with increased productivity but also correlates with higher return on assets and market value.[4]

Collecting and analysing data can help companies make more informed decisions. Retailers, for instance, can replace the "reduce prices and increase sales" strategy with more intelligent marketing campaigns based on collecting and analysing data about the purchasing behaviour of their customers.[3] Marketing in general is a prominent field for data science applications as these principles and techniques can be used for targeted marketing and online advertising as well as for general customer relationship management, such as analysing customer behaviour in order to avoid attrition or to maximise expected value [4].

Provost and Fawcett [4] introduce the idea of big data 1.0 and big data 2.0. They conclude that in the first phase of big data, companies began to build capabilities

to process large data primarily to support their current operations. In this second phase they are asking questions like, what can now be done that was not possible before, or what can be done better? At this point, data science principles and techniques will be applied more broadly and deeply than before. They emphasise that data-analytic thinking will become an important skill throughout organisations, not just for data scientists.

As big data and data analytics become increasingly popular, managers face the challenge of how to leverage them in order to create business value. Creating value from data requires a data-oriented culture and data analytics capabilities inside the organisation. Vidgen et al. [5] studied management challenges in creating value from business analytics. They found that the challenge areas included having a clear data and analytics strategy and the right people to enable a data-driven cultural change. Data and information ethics also needed to be considered when using data for competitive advantages. They emphasise that becoming data-driven is not simply a technical issue. Firstly, organisations need to build business analytics departments with sufficient data analytics capabilities, and secondly these capabilities need to be aligned with business strategy to tackle data analytics challenges in a systematic manner.

The availability of free or low-cost analytics tools along with affordable data storage and computational services offered by cloud providers have also helped small businesses to start taking advantage of data analytics in their daily operations [3]. The term "small and medium sized enterprise" refers to companies that employ less than 250 people and create an annual turnover of less than 50 million or possess a balance sheet total of less than 43 million. In Finland, these companies constitute 99.7% of all companies, employ 65.2% of all employees and account for 59.6% of all value added. The vast majority of Finnish SMEs are micro companies that employ less than 10 people.[6]

SMEs have more limited resources than larger companies and this can have an effect on the way these companies are able to utilise data to support their decision-making processes. For instance, the use of big data analytics in SMEs is lagging far behind in comparison to larger companies. At the same time, the expected global growth rate of SME data analytics market suggests that these companies intent to increasingly take advantage of information-driven decision-making in order to leverage their businesses.[7] Figure 2.1 demonstrates the interaction of profiles involved in the information-driven decision-making process of SMEs. The large number of these companies and their critical role as suppliers in the value chains of larger companies means that providing solutions that will increase the competitiveness of these companies can have a positive effect on the economy as a whole.

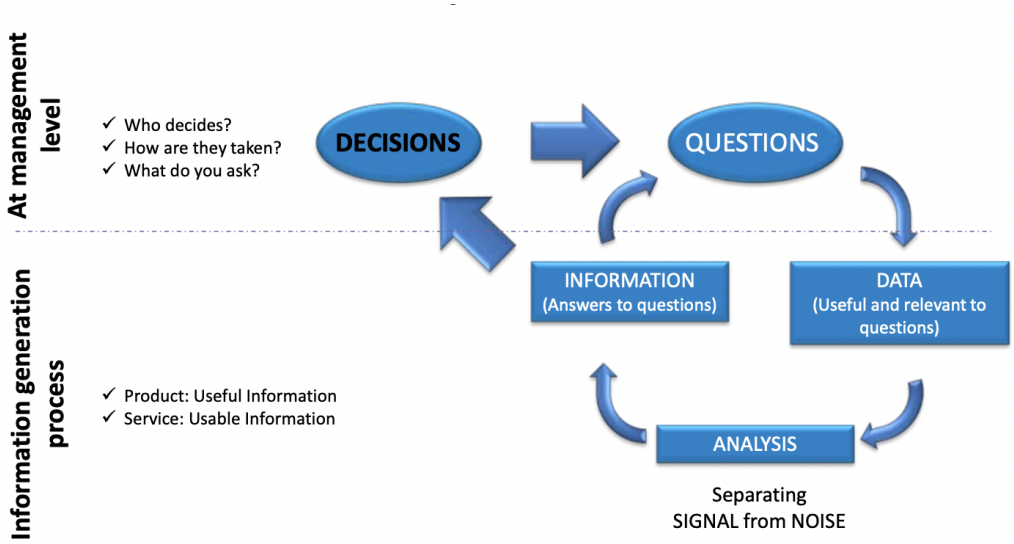


Figure 2.1: Information-driven decision-making in the SME [7]

A growing number of businesses of different sizes seek to utilise data-driven decisions-making in order to improve their performance and competitiveness. Yet, many of them fail to fully take advantage of its possibilities due to the difficulties of aligning their technological solutions with the adoption of data-driven decision-making processes. Maturity models can guide SMEs in their journey towards be-

coming data-driven organisation by increasing company self-knowledge of current state and providing a roadmap for future steps.[8]

Parra et al. [8] studied the usefulness of such a maturity model in evaluating the state of data-driven decision-making in SMEs. In their interviews with family owned SMEs, they found that senior management was very interested in increasing their knowledge of how to better leverage data and adopt analytical practices. Also, the proposed assessment tool proved to be helpful in evaluating how an organisation uses collected data to support its decision-making processes. Their model considers five dimensions: *data availability*, *data quality*, *data analysis and insights*, *information use* and *decision-making*. These dimensions derive from the notion that end users need appropriate access to relevant data and that business decisions should be supported by good quality data. The data has to be transformed into meaningful information that is actively used to support decision-making, leading to better informed decisions under a planned and systematic process.

The model developed by Parra et al. [8] recognises five levels of maturity in terms of data-driven decision-making: (1) uninitiated, (2) awareness, (3) proactive adopting, (4) integral embracement or (5) completely embedded. Each dimension of their model is evaluated according to these five stages of maturity and the overall evaluation is obtained by combining the five dimensions. The results of the evaluation can be visualised with the help of a five vertex radar chart and the evaluation process can be repeated every year or two in order to follow progress. Figure 2.2 shows an example of a radar chart used to visualise company goals and accomplishments in the area of data-driven decision-making [7].

2.2 Evolution of data storage systems

The amount of data available and the characteristics of this data have changed during the past decades as a result of technological development. New requirements

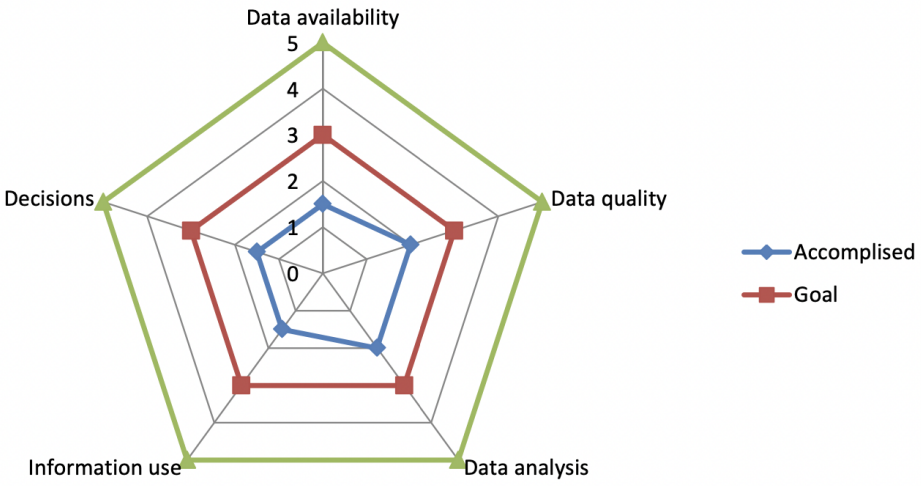


Figure 2.2: Radar chart of information-driven decision-making process [7]

for handling data appear steadily and information storage systems must evolve to meet these needs. In this section we describe types of data and some common data storage systems used in decision support systems with special focus on the differences between the more traditional data warehouse concept and the relatively new data lake concept.

One of the most commonly mentioned modern data themes is the concept of big data as opposed to more conventional IT. Big data is often described through a varying number of adjectives beginning with the letter "V". The original three characteristics of big data are volume, velocity and variety. The list has been later on supplemented with such attributes as veracity, validity, variability, value and visibility.[9] [10] In relation to big data, *volume* refers to the quantity of the data, *velocity* to the speed of data generation or collection and *variety* to the differences in data types and formats. Compared to conventional IT, with big data, huge amounts of different types of data can be streamed in near real-time. *Veracity* refers to the quality of the data - we might be handling precise data such as shopping data or uncertain and imprecise data like medical diagnoses or sensor readings. Closely linked

to veracity is *validity*, which can be used to characterise data interpretations, especially the accuracy and correctness of the data for its intended use case. *Variability* can be used to assess inconsistencies in data and its collection methods. *Value* refers to the usefulness of the data and *visibility* to knowledge visualisation. Optimally, we should be able to attain valuable information and useful knowledge from big data, and visual analysis plays a role in revealing, explaining and discovering these insights. [10]

Businesses collect data from multiple sources such as sensors, social media, web traffic, emails, log files, business applications etc. Data coming from these various sources takes different forms. It can be *structured data*, such as spreadsheets and SQL query results from a relational database or *unstructured data*, like social media posts, emails or satellite imagery. It might also be something in between, *semi-structured*, such as JSON formatted log files or xml data. Structured data has a rigid structure, typically a table structure imposed by an operational or a transactional database system or a data warehouse. Unstructured data on the contrary has no structure. Semi-structured data, on the other hand, can have a nested structure and may not be uniform, meaning that similar records may have a different set of attributes. *Opaque structured data* refers to data that seems to have a certain structure or formal pattern but in reality the schema is not defined making the structure opaque.[11]

Data-driven decisions are typically based on more than one source of data. According to some estimates, five distinct data sources are required on average to reach a data-driven decision. These sources might include both internal and external sources and the chances are that these sources are not of the same format. It has been suggested that 80% of all data world wide could be unstructured by 2025. Therefore, any data-driven decision is likely to be based on a set of sources that include a considerable amount of unstructured data.[12]

2.2.1 From databases to decision support systems

Databases have been used for storing data since the 1950s and relational databases became popular around the 1980s. They can be used to monitor and update real-time structured data.[12] Most enterprise software applications rely on a relational database as the data layer. Multiple autonomous database systems could also be mapped into a single federated database system that has some kind of a meta-database management system to bind the individual database management systems and to act as a single interface for external systems such as applications and query engines [11]. While the current market is still dominated by relational databases, NoSQL data management systems are attracting more and more interest as the traditional schema-on-write approaches like extract, transform, load (ETL) processes have proved inefficient for managing large amounts of semi-structured and unstructured data. The schema-on-read approach of NoSQL data management systems does not require defining the structure of the data for storage but only for further analyses and processing.[13]

Traditional database systems can support the daily operations of an organisation but they rarely satisfy all the requirements of data analysis. Operational databases provide transaction processing, concurrency control and recovery techniques to ensure fast access to data and guaranteed data consistency in a multi-user environment. However, typical operational databases do not include historical data and perform poorly when aggregating large volumes of data or executing complex queries that require joining many relational tables.[14] Furthermore, companies might have several operational systems, each with their own data repository. Data warehouses were developed to support the flow of data from these operational systems into decision-making systems. A data warehouse can hold records from various systems and time frames, and the data can be optimised for fast analyses.

A data warehouse is a type of decision support system that is designed especially

for data analysis and business intelligence and typically allows the users to explore and navigate the data at different levels of granularity [11]. It is a tool that provides analysts fast access to large sets of aggregated data integrated from various sources [15]. According to Malinowski et al. [14] data warehouses contain *subject oriented*, *integrated*, *nonvolatile* and *time-varying* data used to support managerial decision making. This means that they are oriented towards a particular subject of analysis and attempt to integrate relevant data from several operational systems. They do not allow modifications and removals of data, but offer the possibility to retain different values for the same information. Data warehouses can be centralised, encompassing all functional and departmental areas of an organisation, as in the case of *an enterprise data warehouse*, or smaller and more specialised, targeting a particular functional area or user group inside an organisation, like in the case of *a data mart*.[14]

Fast growing, extensive and diverse data can be too big for traditional data storage solutions, too unstructured for traditional business intelligence tools and too fast for traditional static data warehouses.[3] The benefit of a data lake over a data warehouse is that data from different sources and with various formats can be stored in a single repository.[12] Data lakes are typically described as single repositories for heterogeneous enterprise data coming from various sources, stored in raw format and consumed for various analytical activities in order to provide business value [16]. They can be used to decouple data producers, such as operational systems from data users, like reporting and analysis systems. Sometimes the operational systems are not owned by the company as in the case of legacy mainframes or modern SaaS applications. In addition, data lakes can be convenient storage layers for experimenting with data science. They typically provide reliable storage together with computational frameworks, such as Hadoop or Apache Spark. Data lakes come with suitable tools for data governance, data discovery, extraction, cleaning and

integration.[17]

2.2.2 From data warehouses to data lakes

Databases and data warehouses require a predefined schema for the data that will be inserted into the data storage system and all the incoming data must match the schema in order to be inserted successfully. These data storage systems are typically designed for a specific purpose making it easier to define the constraints in advance as part of the business process. This schema-on-write approach can support data integrity and correctness and can also help data cleaning and transformations, but at the same time data ingestion becomes slow and handling of dynamic data may not be possible. Data lakes operate on a schema-on-read approach, meaning that the schema for data is defined once the data is processed. This approach enables fast and easy data ingestion and handling, but poses challenges for data analysis, which requires understanding the structure of the data to be analysed.[11]

Data warehouses are typically built on top of existing information systems in order to support the decision making of a certain user community for which there is a sufficient economic benefit to overcome the costs of the investment. Data warehouses are used to answer predetermined questions based on a single data source. Data lakes on the other hand have the potential to answer a large portfolio of possible questions from multiple data sources making them of interest to a wider range of user communities within an organisation.[18]

The structure and the usage of data had to be predefined and fixed in the traditional solutions like data warehouses and data marts, and data ingestion required rigorous data extraction, transformation and cleansing. The data lake with its focus on storing raw data can avoid or delay such expensive standard processing.[13] This is especially relevant when working with unstructured data without a predefined use case in place. While unstructured data could be cleansed and prepared so that

it was possible to store it in a database or a data warehouse, this is very difficult to do without a clear use case in mind. Data lakes might be more suitable data repositories in the fast changing business environment of today. They allow storing vast amounts of data for future use cases that have not been recognised yet and this can be done in a cost effective manner.

Inexpensive storage is such a profound element of data lakes that they are sometimes seen as analogical to big data technologies such as Hadoop and the Hadoop Distributed File System (HDFS). However, this is a very narrow view that undermines the role of other storage systems such as NoSQL databases in building data lakes and the diverse tools landscape available for data lake practitioners to choose the most appropriate tool for each data processing task.[19] Likewise, Hai et al. [13] emphasise that the data lake should not be viewed solely as a raw data storage system but rather it needs to also provide a set of functions such as metadata management to manage and govern data in order for it to be usable for on-demand data processing and querying.

One of the most profound differences between data warehouses and data lakes is that data warehouses are used for storing data that has been cleansed based on predetermined schema, while data lakes take advantage of low cost technologies to ingest all types of raw data in their native format, providing flexibility and scalability [20]. Despite some obvious differences, these two decision support systems can also coexist together and serve each other. Data warehouses can be the source of data lakes, meaning that the once processed data from the data warehouse is ingested into the most suitable zone in the data lake. Alternatively, a data lake can be the source for a data warehouse that is maintained for a specific use case.[21] Some main differences between data warehouses and data lakes are summarised in Table 2.1.

To conclude, businesses today need to handle a variety of data types coming from numerous sources. Data can be hosted on-premises, in the cloud or it might

Table 2.1: Data warehouse versus data lake

Characteristics	Data warehouses	Data lakes
Time-span	Present	Past and present
Schema	On-write	On-read
Purpose	Predetermined	Evolving
Typical users	Managers, business analysts	Data scientists
Use cases	Reporting, BI, visualisations	Machine learning, discovery, profiling etc.
Data sources	Operational or transactional databases	Multiple including databases, sensors, social media etc.
Data ingestion	ETL	Load-as-is
Data accessibility	Complex	Easy
Data format	Structured	Heterogeneous
Data variety	Low	High
Data maturity	Refined	Raw
Data storage	Relational databases	Multiple including Hadoop, relational db, NoSQL etc.
Data access	SQL queries	Query languages and programming languages

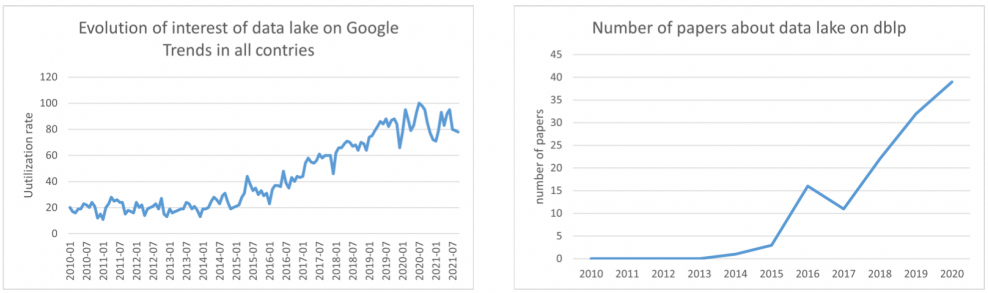
be coming from external sources. Furthermore, the amount of data is growing at a fast pace and it is used for multiple types of workloads and use cases. Modern data lake infrastructures can help data-rich companies manage these challenges. It is common for companies that have prior investment in legacy data warehouses and mainframe technologies to look into more sophisticated data lake infrastructures in order to exploit possibilities brought by more diverse data types and to improve the efficiency of legacy systems by offloading capacity to more flexible solutions.[22] Taking advantage of modern decision support systems such as data lakes can increase data quality inside an organisation. An Aberdeen study [22] showed that satisfaction in terms of data metrics like quality, timeliness and sophistication were higher among those companies that had already invested in data lake technologies compared to similar companies only planning such investment.

The huge amount of heterogeneous data brought by digital transformation is challenging the traditional decision support systems such as data warehouses. At the same time, data lakes enable leveraging advanced data analysis methods such as data mining, text analytics or machine learning. Data lakes appear as a new and attractive alternative for storing data to be used to support decision-making in companies of all sizes. The next section elaborates on the data lake concept and relevant themes such as data lake architecture and data lake management.

2.3 The data lake concept

The data lake concept is a relatively new one - first introduced by Pentaho CTO James Dixon [23] in 2010. He presents the idea of a data lake as a better approach to handling and analysing big data compared to traditional solutions such as data marts. "If you think of a data mart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a

source to fill the lake, and various users of the lake can come to examine, dive in, or take samples." Since then, there has been a growing interest towards this concept among practitioners as well as in academia. Figure 2.3 by Zhao [24] demonstrates the evolution of interest in data lakes. The amount of academic research on data lakes has grown significantly since 2014 as researchers have attempted to define the concept and studied relevant themes like data lake architectures and metadata management.



(a) General evolution (b) Evolution in academic world

Figure 2.3: Evolution of the interest of data lake [24]

Sawadogo and Darmont [21] reviewed earlier literature on data lakes in order to provide a more comprehensive definition of the concept. They built upon the previous attempts to define the concept and arrive at the following definition: *"A data lake is a scalable storage and analysis system for data of any type, retained in their native format and used mainly by data specialists (statisticians, data scientists or analysts) for knowledge extraction."* They complement their definition by listing six characteristics that should be included in data lakes:

1. a metadata catalogue that enforces data quality
2. data governance policies and tools
3. accessibility to various kinds of users
4. integration of any type of data

5. a logical and physical organisation

6. scalability in terms of storage and processing

Similarly, Ravat and Zhao [20] find previous definitions of data lakes too vague and attempt to provide a more complete definition that includes input, process, output and governance of data lakes. They define a data lake as *"a big data analytics solution that ingests heterogeneously structured raw data from various sources (local or external to the organisation) and stores these raw data in their native format, allows to process data according to different requirements and provides accesses of available data to different users (data scientists, data analysts, BI professionals etc.) for statistical analysis, Business Intelligence (BI), Machine Learning (ML) etc., and governs data to insure the data quality, data security and data life-cycle"* [20].

2.3.1 Data lake architectures

Data lake architectures have evolved from simple flat architectures with single data repositories towards more complex multi-ponds and zoned architectures, which also reflect the activities performed in a data lake and the consequent need to have multiple data storage units inside the data lake. The original *flat data lake architecture* allowed the idea of loading heterogeneous, voluminous data in its raw format at a low cost, but was too simplistic and closely tied to the Hadoop system. *Multi-ponds data lake architectures* enabled having multiple storage areas for different structural types of data making finding data faster and analysing it easier.[24] In this model, the data is initially ingested into a *raw data pond* and then transformed and moved to other ponds according to its characteristics, such as *analog pond* for high velocity IoT data or *textual data pond* for unstructured textual data. Valuable data can be secured in a long-term *archival pond*. In multi-pond data lake architectures, the data is always available in a single pond at a given time.[19]

Zoned data lake architectures emerged to overcome some of the limitations of the previous architectural models. In this approach, data is assigned to zones inside the data lake according to the level of processing that has been applied to it. The initial zone is always *a raw data zone* where data is ingested in its raw format. In contrast to multi-pond architectures, the raw data remains available in the raw data zone indefinitely. The ultimate number of suggested zones differs between alternative zoned architectures.[19] In Zaloni's zoned data lake architecture, for instance, the data is assigned to zones based on its maturity. *Transient loading zone* and *raw data zone* are accompanied by a *refined data zone* for data analysis and a *trusted data zone* for cleansed data. Data can be explored by data scientists in the *discovery sandbox zone* and by business users with the help of dashboard tools in the *consumption zone*. *Governance zone* enables managing and monitoring the data lake.[24] [21]

Existing data lake architectures are frequently supported by technical solutions and hence do not offer a clear distinction between functionality-related and technology-related elements of the entity. Ravat and Zhao [20] wanted to distinguish between the two by approaching the subject from a usage perspective. They provide a functional data lake architecture that can be implemented through different technical solutions. Their model is shown in Figure 2.4. This functional data lake architecture suggested by Ravat and Zhao contains four essential zones: a raw data zone, a process zone, an access zone and a govern zone. The first three each include a separate treatment area and a data storage area for storing results from these processes. The *raw data zone* is for all types of data ingested without processing and stored in their native format. The *process zone* is for transforming the data according to users' requirements and storing the intermediate data. The *access zone* is for storing data for analytics and allowing access to self-service data consumption such as reports, statistical analysis, business intelligence analysis and machine learning algorithms. The fourth zone is applied to each of the other zones. This *governance zone* is

for ensuring data security, data quality, data life-cycle, data access and metadata management.[20]

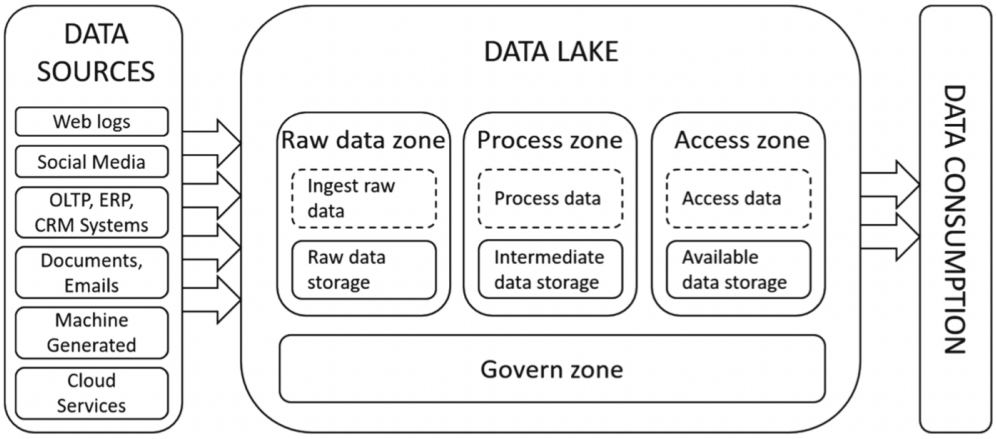


Figure 2.4: Data lake functional architecture [20]

Many data lake architectures suggested in literature focus on the characteristics of data and the transformation processes applied to this data, while users of the data lake receive less attention. Hai et al. [13] note that people interact with the data lake in different roles. They describe four types of user roles that are present in a typical business data lake scenario: (1) data scientists and business analysts, (2) information curators, (3) the governance, risk, and compliance team, and (4) an operations team. The first group builds and applies analytics models over the data lake while the second one defines new data sources and maintains metadata over the existing data sources of the data lake. The third group ensures that organisational regulations and business policies are followed. Finally, the fourth group is in charge of maintaining the data lake and may include specialists such as data quality analysts and integration developers.

2.3.2 Data lake management

Building a data lake typically involves several stages such as data ingestion, data processing and transformation, data analytics and visualizations.[16] Data ingestion refers to the process through which data is loaded into the data lake and stored in databases or file systems. This might require a sophisticated scheduling approach in order to synchronise multiple incoming data streams.[13] Data ingestion can be done in batches or real-time via streaming, and both of these modes can be used in a given data lake for different types of data. Batch processing is an efficient way of processing large volumes of non time-critical data. The steps of gathering and extracting data, as well as processing, enriching and formatting this data to fit its intended use case, can be automated via batch processing. Real-time processing can be used to process an unbounded stream of data that requires a low response time such as IoT sensor data that is used to generate alerts.[16]

Data lakes are dynamic in nature meaning that new datasets as well as new versions of existing files are steadily pouring into the data lake through the ingestion process. Providing efficient and cost-effective storage and retrieval of versioned data is a critical feature of data lakes. This can be especially challenging when there are peculiarities of data formats, which require schema evolution between versions.[17] The data ingestion phase must also involve extracting and storing metadata in addition to the actual data in order to ensure the usability of the data later on [24].

A critical question in building a data lake is, what kind of a storage system should be used to preserve the ingested data. The storage layer could include relational or NoSQL databases or some kind of a combination of these. Systems that provide integrated access to a configuration of multiple data stores for heterogeneous data are called polystores. The Hadoop Distributed File System (HDFS) is often suggested as the data lake storage solution. The HDFS supports multiple text file formats like CSV, XML and JSON as well as binary files like images. Data can be compressed

with formats such as Snappy and Gzip and stored in columnar or row-based formats like Parquet or Avro. Cloud providers have blob storage solutions, such as Azure Blob Storage by Microsoft and S3 by Amazon Web Services, that are optimised for large, unstructured object data.[13]

Having raw data in a data lake does not automatically mean it can be processed and analysed easily. Rather it requires knowledge on the location and the structure of the data as well as transformation into a format that supports the given use case. In most cases the raw data in the data lake needs to be processed before it can be used for analysis. This can be a time consuming endeavour - according to some estimates, only 30% of data life cycle is spent on actual data analysis, while 70% is spent on finding, interpreting, cleaning and integrating data [11]. Data extraction refers to transforming the raw data to a predetermined data model, often an integral part of preparing for data discovery.[17]

Data transformation in a data lake can include extract, load and transform (ETL) jobs, typically associated with data ingestion in a data warehouse system. These jobs are used to extract relevant data from various sources (extract), converting it into coherent, reliable data for decision making (transform) and inserting it into a permanent data storage system (load).[24] The data transformation process may also include tasks like data cleaning, data integration and data enrichment. Machine learning algorithms can be used to perform data exploration, classification, and prediction.[11] The transformed data is stored back into the data lake, meaning that the data lake includes different versions of the same data at varying levels of refinement.

Data lakes need to be managed in order to avoid them turning into messy, unmanageable *data swamps*, which is a term used for a deteriorated data lake. A data swamp is an unmanaged data lake that has become inaccessible to its intended users or simply does not provide any value to them.[10] Data lakes are complex systems

that need careful management. They contain various types of data that is ingested through different data transfer modes without a predefined schema and frequently without an explicit use case in mind. Furthermore, the data can be transformed into numerous datasets that are analysed by users with varying specialisations and interests. These characteristics of data lakes highlight the need for effective governance in order to avoid the data lake from turning into a data swamp.[24] Taniar et al. [11] emphasise the need for a clear strategy or design methodology for data lakes in order to transform the promises of this technology into reality.

Hai et al. [13] identify six types of activities related to data lake maintenance. *Dataset organisation* refers to solving the problems of structuring and navigating large, heterogeneous datasets so that the users of the data lake can effectively find their desired datasets. It often requires profiling datasets with their metadata. *Related dataset discovery* refers to finding a subset of relevant datasets that are similar or complementary to a given dataset. *Data integration* refers to providing the users with a unified data access to a combination of various heterogeneous data sources via such data integration techniques as schema matching, schema mapping, query reformulation, entity linkage, etc. *Metadata enrichment* refers to the steps performed to discover additional dataset metadata such as semantic knowledge, relationships with other datasets or constraints to be used while cleaning the data. *Data quality management* is used to guarantee quality and efficiency of query results. Data quality can be improved by obtaining dependencies from the data and identifying objects that violate these dependencies as potentially erroneous data. *Schema evolution* is needed to handle changes in the schema and integrity constraint of incoming data compared to existing data in the data lake.[13]

Metadata management has emerged as an important concept in the literature on data lakes. Zhao [24] defines metadata in the context of data lakes as *"structured information to describe and explain all the resources (datasets, preparation,*

analyses) stored in the data lake, and relationships between them as well as common sense about them". Dataset metadata can be further classified into three categories based on its characteristics. *Intra-metadata* is dataset specific information, including *dataset properties* like name and creation date, *schematic metadata* used to describe data structure like attributes and property information, and *basic semantic metadata* that describes dataset meaning like tags and descriptions. *Inter-metadata* is used to describe relationships and similarity/dissimilarity between datasets, while *global metadata* is concerned with advanced semantic annotations and exists independently from any datasets.[24]

Metadata management refers to the process through which the availability and quality of metadata is ensured, safeguarding effective usage of the stored elements in the data lake. Metadata management systems can be used to automate the generation of metadata and providing user interfaces for exploring metadata.[24] Multiple metadata management models and systems have been suggested and reviewed in the context of data lakes but this vast topic is outside of the scope of this study. Instead, the role of cloud computing is discussed next, since the availability of low cost storage combined with data processing and analysis tools has an important enabling role for small businesses attempting to take advantage of data in their decision making processes.

2.4 The role of cloud computing

Gupta et al. [25] studied the adoption of cloud services in small businesses. Based on a literature review, they identify several potential motivators that could drive SMEs toward the adoption of cloud computing services. These include cost reductions, ease of use and convenience, reliability, sharing and collaboration as well as security and privacy. They suggest that the adoption of cloud computing services can bring cost reductions to SMEs as they only pay for the resources they use without

upfront investment, shifting the nature of IT costs from capital costs to operative costs. Also, sharing and collaboration can be easier when resources are delivered online. Furthermore, these resources are accessible anytime and anywhere, which makes them more reliable, and the businesses that use them more productive and efficient. In their empirical data, the ease of use and convenience appeared as the most significant motivator for SME cloud adoption, followed by security and privacy and lastly cost reduction.

Cloud computing comprises three types of services: Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS). The biggest and most mature type of cloud services is the SaaS model, in which applications like enterprise resource planning (ERP) and customer relationship management (CRM) systems are made available through the Internet, instead of installing them on the end user's computer. In the case of the PaaS model, consumers purchase platform resources, like operating systems and databases, or tools, like Java or .NET, from commercial vendors such as Amazon Web Services over the Internet, instead of acquiring licences directly. Similarly, IT infrastructure, like servers and storage devices, can be purchased from cloud vendors, who maintain the physical devices in their data centres, and provide access to these resources through the Internet. Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3) are well known examples of such IaaS services.[25]

Early literature on data lakes often viewed the concept as equivalent to using free or low-cost technologies, such as Hadoop or Apache Spark, for storing, processing and exploring raw data. Later on, this view has been challenged as the concept of data lake is now also linked to proprietary cloud solutions, such as Azure and Amazon Web Services.[21] Commercial data lakes are often hosted on cloud platforms, since they offer many advantages for building data lakes like scalability, costs savings and support for analytics. Data storage and computation power can

be scaled dynamically in a cloud data lake. The data lake owner pays only for the resources that are used and the cost of running these resources tends to be lower on a cloud platform compared to hosting on-premises. Cloud platforms also provide useful data visualisation and analytics tools, such as machine learning services, that can be used for gaining insights from the data stored in a data lake.[13]

Analytical projects like data lakes often do not have clearly defined targets. It might be unclear in the beginning just how much data will be ingested, where it will come from and what kind of storage, computational capacity and services are needed to process the data in order to deliver business value. Such a project can be more of an exploratory journey with aspirational business outcomes. Cloud services are a prominent option with these types of projects, since they enable getting started easily and running experiments. Building a business data lake on a cloud platform such as Amazon Web Services should start with stating business goals that can support the business strategy of the company. Next, a set of measurable business outcomes that would have a positive effect on these goals need to be identified, along with metrics that can be measured in order to validate the success of the project. It is a good idea to run small experiments to validate business value before building a more comprehensive solution. Experiments are a means to assess technologies and to explore data, and they can indicate whether a solution can deliver business value and how the solution should be designed to realise this value.[16]

There are many opportunities for lowering costs by making the right decisions when building a cloud based data lake. The costs of operating a cloud based data lake can be reduced by using data compression, reducing run frequency, choosing a columnar data format, creating data life-cycle policies and opting for serverless services. *Data compression* reduces the costs of storing and scanning the data. *Reducing run frequency* to match the minimum data timeliness requirements instead of constantly transferring all available data will reduce costs by cutting the total

run time of managed cloud services. *Partitioning data* will effectively cut the costs and improve performance of queries by reducing the amount of data that is scanned. *Choosing a columnar data format* such as parquet will have similar effects via reducing disk I/O requirements. Reducing costs via *data life-cycle policies* refers to moving data between storage tiers - raw data can be archived in the cheapest storage tier once it has been processed and is no longer needed at daily bases, but must be kept in case it needs to be reprocessed as new requirements surface. *Opting for managed services* can reduce costs since serverless services only incur costs when they are used.[16]

3 Case study: Greenlips Beauty

3.1 Greenlips Finland Oy

Greenlips Finland Oy is a small, Finnish company operating in the import and sales of natural cosmetics produced in Europe. It was established by its CEO, Sharon Forbes, in 2010 and is best known under the trade name of Greenlips Beauty. The company had four employees and an annual turnover of 607 000 euros in 2021. It is located in Turku, Finland, but has a strong online presence. Greenlips Finland Oy has a beauty salon in the centre of Turku. Greenlips Beauty is one of the pioneers in the natural cosmetics business in Finland. The cosmetics industry worldwide is a large business sector that does not traditionally operate in the most ethical and sustainable ways. The people at Greenlips Beauty work towards a more ethical and sustainable beauty industry by offering consumers the choice of natural cosmetics over traditional cosmetics. One of their core values is respect towards people, animals and nature.[26]

Greenlips Beauty imports carefully selected European cosmetics brands. At the moment, they represent 6 different brands: Abel, Kaerel, Less Is More, Petit&Jolie, Und Gretel and Uoga Uoga. These brands offer a wide variety of products, including scents, skin care products for women, men and children, and professional hair and make-up products. To have a closer look at their products and services, visit Greenlips Beauty online Store at `https://greenlipsbeauty.com`. These products

are sold to both, business clients and consumers, with business clients representing over 85% of the total sales value in 2021. Business clients are typically beauty professionals who purchase raw materials for providing beauty services or resellers.

Interviews with the CEO of Greenlips Finland Oy reveal that maintaining a network of sales representatives is an important part of the company's everyday operations. Greenlips Beauty offers its advanced product and sales knowledge to its sales representatives in order to help them succeed in their business. This contributes to higher sales also for Greenlips Beauty and leads to wider availability of safe and sustainable local beauty services throughout Finland. Furthermore, the CEO of the company states that they are interested in finding new business opportunities in this type of mutually beneficial cooperation with their sales representatives and other actors in the sustainable beauty industry. Many specialists in this area have little formal education on business administration and marketing activities. Preliminary experiments support the notion that mentoring can have a positive effect on their business performance and there might be a market for such mentoring services.

3.2 Technical business environment and limitations

Like most small businesses nowadays, Greenlips Finland Oy is using SaaS based business software applications to sell its products, manage its customer relationships and to handle its accounting, among other things. The majority of sales come through their online store, which at the moment is run on the MyCashflow e-commerce platform. Their beauty salon uses Timma Pro - a salon management tool that is specifically designed for hair, beauty and wellness professionals. Their sales and marketing team is using Monday.com for keeping track of on-going marketing tasks and planning for future tasks together as a team. All accounting related tasks have been outsourced to an independent accountant who uses Procountor accounting software. The benefits and limitations of these third-party applications are

described in more detail in the next sections.

3.2.1 MyCashflow e-commerce platform

MyCashflow is an e-commerce platform developed by a Finnish company Pulse247 Oy. The company was established in Kajaani in 2007 and has over 30 employees. Its MyCashflow e-commerce platform is used by over 2000 companies which makes it one of the most used e-commerce platforms in Finland [27]. MyCashflow software offers a wide range of features for running an online store. It is a SaaS product that has four subscription levels: Basic, Advanced, Pro and Enterprise. The Advanced, Pro and Enterprise level subscriptions offer more features than the Basic level subscription for a greater monthly fee. Greenlips Beauty has settled for Advanced subscription level. This level includes a Rest API, but has a more limited number of products, storage and shop language versions available compared to the Pro level subscription or the Enterprise level subscription, which offers a customised e-commerce solution. The MyCashflow Rest API can be used to integrate an online store with a number of third-party applications such as CRM and accounting software. The MyCashflow Rest API has an online API reference available at `https://support.mycashflow.com/api`.

Selecting the most suitable e-commerce platform is a critical decision for any sales organisation that accumulates the majority of its sales online. Greenlips Beauty chose MyCashflow based on its previous experiences with e-commerce platforms and a review of other possible solutions. They decided to go for a managed solution instead of a customised, self-hosted solution such as Magento, which they were using prior to MyCashflow. Maintaining Magento was expensive, since it required having an external IT specialist to perform maintenance tasks and updates. Despite this support, many problems prevailed such as regular downtime, which went from hours per month to only minutes per month, with the adoption of the lighter and more

cost-effective MyCashflow. With this platform, support is included in the price and the software is developed by the service provider to cover the most commonly requested features. Settling for a managed solution came with some drawbacks related to reduced adaptability and aesthetics as well as the loss of a customised customer loyalty program. However, in terms of data-driven decision-making, the most critical limitations of this system have to do with reporting.

MyCashflow offers some off-the-shelf reports such as *the customer report* that allows reviewing the purchasing behaviour of registered customers and could be used to reward loyal customers or to lure back inactive customers. *The order report* allows reviewing a list of incoming orders while *the product sales report* shows sales by product for a specific time frame. Users can narrow down these off-the-shelf reports by setting time frames and filtering conditions and the reports can be exported as CSV files. Yet, they do not provide aggregated data necessary to answer such questions as, what are the most trending hair care products among the top resellers in the Helsinki area? Users can also create customised reports, which increases the usability of the data, but requires added manual labour. For instance, if creating a monthly customised customer report would take approximately 5 minutes per customer, this would mean that simply creating reports for the 100 most important customers would drain a whole working day and the data would still be limited and vulnerable to selection bias.

3.2.2 Salon management with Timma Pro

Timma Pro is a salon management tool designed for hair, beauty and wellness professionals. It is used by 13 000 professionals in the Nordics. It is designed to be a seamless solution from calendar to cashier and marketing. It can be used to manage bookings, to receive payments, to create a promotional web page and to communicate with customers. The platform includes an online booking system,

where customers can make reservations any time of the day with ease.[28] Timma Pro offers some reporting and statistics features and supports data exports in CSV format, but does not have an application programming interface, which limits the possibilities for integrations and automatic data transfer.

Timma Pro was chosen for Greenlips Beauty, because it allows customers to make bookings without registration. This was such a critical feature that it justified abandoning their previous cashier and online reservation system, which had more sophisticated reporting features. In terms of data-driven decision-making, the current system provides information about the utilisation rate of the salon but lacks, for instance, support for customer segmentation based on the frequency of visits and metrics such as average service and product purchases in euros per customer.

3.2.3 Customer Relationship Management with Monday.com

Monday.com is a versatile team working software, which can be used for managing many types of workflows. It can be used for product management and managing sales and marketing activities or as a customer relationship management solution, among other things. Monday.com users create boards to keep track of and manage their projects. A board is a kind of a virtual white board that is interactive and allows easy collaboration between team members.[29] Monday.com has an application programming interface making it easily integrated to other systems. The Monday.com GraphQL API has an online API reference available at `https://developer.monday.com/api-reference/docs`.

Greenlips Beauty uses Monday.com for business management in general and for customer relationship management in particular. It is a kind of a central repository for managing any business and marketing activities that require planning and team working, such as product launches, marketing campaigns and organising events. It is used for storing data that needs to be kept for reference and shared by the team

such as instructions and notes or links to other systems. One of the main benefits of this system is that it enables team working in a flexible way. It is easy to assign responsibility and share notes between team members, and the data is secured and accessible at all times. This type of setup is essential for growth, especially in a work community that favours remote work.

The challenges with Monday.com, and customer relationship management in general, relate to designing optimal workflows and committing people to these processes to ensure data is in fact recorded and up-to-date. This is also a critical requirement for data-driven decision-making, since data must be accumulated in a systematic way in order to be able to collect and analyse it later on. The case company is struggling with having all the relevant customer data in the system and keeping it updated manually. There is a clear need for software integration between MyCashflow and Monday.com in order to automate the flow of customer and sales data from the web shop into the CRM system. Manually adding customer data to Monday.com is slow and error prone. Monday.com supports importing and exporting data via excel files, which could be used to transfer data between systems, but it is not possible to update existing boards in this way. Furthermore, the key performance indicators (KPIs) needed for planning customer relationships and marketing activities are not readily available in the source systems. According to management's assessment, calculating and updating this data manually takes approximately 8 minutes per customer meaning that keeping data updated for the whole customer base would be a full time job.

3.2.4 Procountor accounting software

Procountor is a popular accounting software used by independent entrepreneurs as well as established companies offering accounting and financial services. Procountor includes all necessary features for fully digital book keeping. The company also

provides related services like a digital signature service and factoring services. Procountor has an application programming interface that enables integrations between Procountor and other business software.[30]

Greenlips Beauty is using Procountor since it is the software provided by their external accountant. They have an existing integration in place between MyCashflow and Procountor that enables transferring invoices from the web shop into the accounting software via a button click in the web shop in order to handle accounts receivable in the accounting software. Procountor contains detailed data of company costs and provides reporting of financial performance for public and internal use cases. It enables monitoring selected company operations with the help of dimensions and could be used to budget future costs and revenues in addition to recording actualized transactions. At the moment, the case company is mostly interested in utilising sales data in their decision-making processes, but accounting data could be used in the future to analyse the roots of divergence between targeted and actual performance as well as the relationships between sales and marketing activities and the overall financial performance of the company.

3.3 Business objectives and technical requirements

The management of the case company seeks to move from making intuitive business decisions towards a more data-driven decision-making model. In order to succeed in this aim, they need better visibility to their data. The case company operates in sales, emphasising the role of sales and marketing data in the decision-making process. In the beginning of the project, they have some concrete objectives related to sales reporting and data visualisation, as well as more vague objectives concerning combining data from various sources in order to assess the effectiveness of their marketing activities, and even more aspirational outcomes such as identifying new business opportunities. The management is relatively content with the functional

features of their existing SaaS applications and is not looking to invest in new systems, but rather find a solution that tackles the reporting issues of their existing systems while providing tools for data analytics and visualisation.

The majority of sales are accumulated via the MyCashflow web shop making this the most important source of sales data. The reporting provided by this e-commerce platform is too simple to answer complicated business questions and requires too much manual labour to be cost-effectively used on a daily basis. The management needs to receive weekly sales reports and visualisations that allow examining the evolution of sales during the past years by customer group, brand and individual customer. They need to be able to combine and aggregate customer, order and product data in order to identify trends and segments and to make calculations and comparisons.

The management needs to incorporate sales data to their everyday activities without costly and error prone manual labour. This requires integrations between distributed systems: sales KPI data needs to flow from the web shop into the CRM solution automatically and in a timely manner. Access to accurate and timely data in itself promotes data-driven decision-making by provoking thought processes and observation of cause and effect. At the same time, not having timely information in the CRM system, and thus failing to respond to this information appropriately and on time, can have negative consequences ranging from a minor loss of sales to the termination of a lucrative business relationship. Furthermore, it can be difficult to assess such indirect ramifications as a friction in an existing customer relationships or a missed business opportunity in strictly financial terms.

The management of the case company wants to be able to assess the effectiveness of marketing efforts and eventually allocate limited resources, financial and temporal, to activities that yield the highest return, be it in terms of customer satisfaction, brand awareness, increasing profits or something else. Marketing activities often

take place in social media meaning there should be a way to connect marketing activities such as social media campaigns to concrete sales transactions. The solution should enable collecting data on sales and marketing activities and assessing their effectiveness systematically and in a reproducible way in terms of resulting sales.

With these objectives and limitations in mind, a cloud based data lake appears as a prominent solution. It provides a platform for ingesting and storing data from various systems at a low cost. It provides tools for integrating and transforming data to gain meaningful insights. The processed data can be visualised with BI tools or displayed inside the existing systems via automated data transfers. It will remove the need for manual labour and ensure data accuracy and timeliness. Implementing the data lake requires data engineering skills, but the solution should not require extensive maintenance once installed and running. Connecting the data lake with Microsoft Power BI data visualisation tool offers management of the case company flexible visibility to the transformed data and allows them to build visualisations and dashboards according to their existing and future needs without extensive knowledge of SQL or programming languages. This approach is consistent with the learning by doing mentality characteristic for entrepreneurs and the management of small businesses.

4 Research methodology

4.1 Application of design science principles

In this thesis we aim to follow the seven research guidelines proposed by Hevner et al. [2] in their essay on design science in information systems research. First of all, this research method should produce a viable IT artefact, which could be a construct, a model, a method or an instantiation. These artefacts are interdependent with the organisational and social context of their users. Secondly, the research process should address a relevant organisational problem. The role of the artefact is to demonstrate the feasibility of the design process and the end product. Information system research in general aims at increasing knowledge and understanding needed to develop and implement technological solutions to unsolved, important business problems. The difference between the current state and the aspired state of a system constitute the business problem to be solved. Business problems are typically related to increasing revenues or cutting costs by adopting effective business processes and information systems have an important enabling role in achieving such goals.

The third guideline emphasises the need to demonstrate the utility of the IT artefact via well-executed evaluation methods. It can be evaluated in terms of its functionality, performance, reliability and usability or any other relevant quality attributes dictated by the business environment. The artefact should also integrate well with the technical infrastructure of the business environment. The evaluation

process provides necessary feedback to the iterative design process that completes only once the solution meets its requirements. Fourthly, design science research should contribute either in the area of the design artefact itself, foundations or methodologies. Most often the contribution is the artefact that provides a solution to an unsolved problem by extending knowledge or applying existing knowledge more innovatively.

The fifth guideline addresses the way the research is conducted, calling for effective use of theoretical foundations and research methodologies. Rigorous methods should be used in both, the construction, and the evaluation of the design artefact. Sixth, the design process is seen as a search process for an effective solution. All possible means that satisfy an end condition under the given circumstances constitute the set of possible solutions to choose from. Lastly, the research must be communicated in sufficient detail. For technical-oriented audiences this means being able to implement the described artefact in a suitable context. For management-oriented audiences it means being able to assess whether the resources of the organisation should be used to implement such an artefact.

During this study we produced a viable IT artefact (as instantiation) that is a data lake instance, built on the cloud computing platform provided by Amazon Web Services. This artefact was built for a case company according to their specific needs. Throughout the process, we paid special attention to the organisational and social context of the case company in order to better understand the business problem at hand and to produce a viable technical solution to this problem. The management of the case company has noticed that their existing software systems do not provide them with enough data for meaningful insights. They want to be able to reliably and effectively assess the impact of efforts such as sales and marketing campaigns on the productivity and profitability of the company. Building a data lake can address this business problem by offering a central storage for all relevant company data

along with tools for data visualisation and analysis.

Our data lake instance demonstrates the feasibility of such cloud based data lakes for SMEs that need better visibility to their data in order to make more solid decisions. The utility of the data lake solution is evaluated according to several criteria: availability, reliability, timeliness and usability of the data. The data lake should provide access to a sufficient amount of data. Users need to be able to connect to the data online and it must be integrated inside the tools used by the marketing team. All relevant data must be included and it must be accurate and frequently updated. The reliability of the data was evaluated at every state during the process. Technical verification was done to make sure that all relevant data was transferred. Semantic verification was done with the company to make sure the data was correct. Data from various sources must be integrated and transformed to a format that enables useful analysis and visualisation and ultimately insight. The solution must provide analytical tools that can be used independently by the management and marketing team in order to be usable.

Data-driven decision-making is not a new phenomenon although its significance for business success is highlighted today as more and more data is available and companies are expected to use this data to become more profitable. The concept of a data lake was first introduced in 2010 and many commercial solutions have emerged since. The novelty of this study derives from the specific research context. We adopt concepts and solutions typically linked to bigger companies and use them in a small business context. This study contributes to the existing research on data-driven decision-making by discussing some of the challenges and benefits from the perspective of a small, sales oriented organisation. We contribute to the existing literature on data lakes by providing a practical example and evaluating its applicability in a less typical setting. This thesis describes the solution from a technical and a business point of view. We give a detailed description of the resulting IT

artefact in Chapter 5 along with a discussion on the achieved business benefits.

4.2 Iterative design process

The design process was iterative, consisting of five stages that included several interviews and workshop sessions with the case company. The process started with a discussion on the preliminary objectives of the company related to data-driven decision-making and the limitations of their current technical environment in terms of reaching these objectives. A cloud based data lake was chosen as the technical solution based on these discussions and a review of possible alternative solutions. During the process, regular interview and demo sessions were conducted with the case company in order to produce an artefact designed to fit the organisational and social context of its users. To better understand the business problem faced by the case company, we discussed the current state and the aspired state of data-driven decision-making in the organisation. In developing the solution we tried to eliminate the differences between these states.

The first stage was dedicated to building the data lake. This stage included choosing a cloud computing platform, identifying the most critical sources of data, setting up storage and data transfer from source systems into the storage layer and configuring the necessary service components to automatically define schemas and create/update data catalogues. The second iteration was dedicated to data transformations and visualisations. The raw data was transformed in an ETL process into a suitable format for visualisation. We used the native business intelligence service offered by AWS, Amazon Quicksight, to implement some of the visualisations specified by the company and to demonstrate the possibilities of data visualisations in general. The third stage was dedicated to integrations needed to reduce manual data transfer between systems. In this stage we added some new data sources and gave the case company management independent access to the data lake through

Microsoft Power BI. The fourth iteration was dedicated to shaping marketing workflows to better support data gathering about these activities and automating data transfers into the data lake. In the fifth iteration we built an interactive Power BI sales dashboard. One of the objectives of this stage was to provide the management with sufficient skills to independently continue creating reports and visualisations with the data lake.

5 Data lake for Greenlips Beauty

5.1 Data lake architecture

We decided to build our cloud based data lake on the cloud computing platform provided by Amazon Web Services. This platform provides a flexible storage system, serverless data processing services and useful tools for data visualisation and analytics at an affordable rate suitable for small businesses. Three similar environments were set up for different purposes: (1) a development environment for developing data ingestion and transformation jobs, (2) a staging environment for testing, performance measurement and quality assurance and (3) a production environment to be used by the management of the case company. Each environment includes the same Amazon data lake components and operates roughly the same way.

Amazon Simple Storage Service (S3) is an object storage service that can be used for storing different types of data, both structured and unstructured. Some of the frequent use cases of S3 include backup, archiving, disaster recovery and cloud-native application data. This flexible storage service is also the foundation of most AWS data lakes. It works seamlessly with a number of native AWS services designed for gaining insight from data, but can also be used with preferred services through integrations with third-party service providers. Data can be protected through identity and access management (IAM) policies at object level and stored at cost-effective storage levels designed for specific use cases, such as archiving. The starting price

of standard S3 storage is $0.023 per GB at the time of writing.

Our data lake content is stored in Amazon S3 with each environment having separate buckets dedicated to different types of data. Data is assigned to buckets based on its level of refinement, following the zoned data lake architecture approach. The *raw data zone* is the immediate storage of data being ingested into the data lake from source systems. Only minimal quality checks and modifications are performed during this initial ingestion process. The raw data is then transformed into a more suitable format for data visualisation and analysis via ETL jobs. The final outputs produced by these ETL processes are compressed and stored in the *transformed data zone*. In addition to these two zones, we have a *query result zone* that is used to store result sets from data queries against the transformed data. The data lake architecture is described in Figure 5.1.

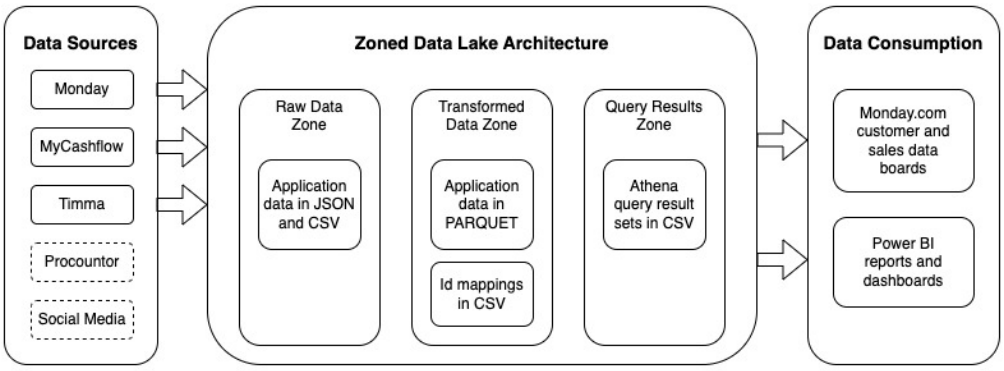


Figure 5.1: Zoned data lake architecture

Some data lake architecture models suggest having an archived zone for data that is accessed more rarely. The Amazon S3 pricing model supports this type of division by offering lower costs for storage classes with infrequent access and longer retrieval periods. The total amount of data stored in our data lake is relatively small, however, so there was no need for a separate archived zone for cost or performance reasons. Similarly, some data lake architecture models have a separate batch processing zone for bulk data and a real-time processing zone for fast data. Since batch processing is

more cost-effective and our targeted time span for analysis allows for weekly batch ingestion instead of real-time streaming, we opted for using only one ingestion mode and having only one ingestion area in the data lake. The archived zone and a real-time streaming zone could be implemented in the future, if the amount of data stored in the data lake increases or the time-span requirements change.

In an attempt to keep the data lake well organised, we issued separate areas inside each data lake zone for data from different source systems. Therefore, application data from MyCashflow e-commerce platform, Monday.com CRM system and Timma Pro salon management system are stored in dedicated folders inside the raw data zone. Alternatively, the zones could have been divided into separate areas based on the format of the stored data. With this type of structure, the JSON formatted application data from MyCashflow and Monday.com would have been stored inside the same folder in the raw data zone. Since the number of source systems is limited in our data lake solution and each system has its own data ingestion and transformation logic, we opted for storing application data separately.

5.2 Steps from data ingestion to visualisation

Transforming raw data into a suitable format for data visualisation and analysis involves multiple consecutive steps, such as data filtering and data type transformation, schema evolution and data catalogue updates as well as data format optimisation. On the AWS cloud computing platform this transformation process can be handled effectively with its serverless data integration service AWS Glue. The next subsections are dedicated to describing the steps involved in the data ingestion, transformation, query and visualisation process in more detail. We use multiple AWS Glue services to automate and orchestrate the process. Figure 5.2 visualises the sequential flow of data between the source systems and these services.

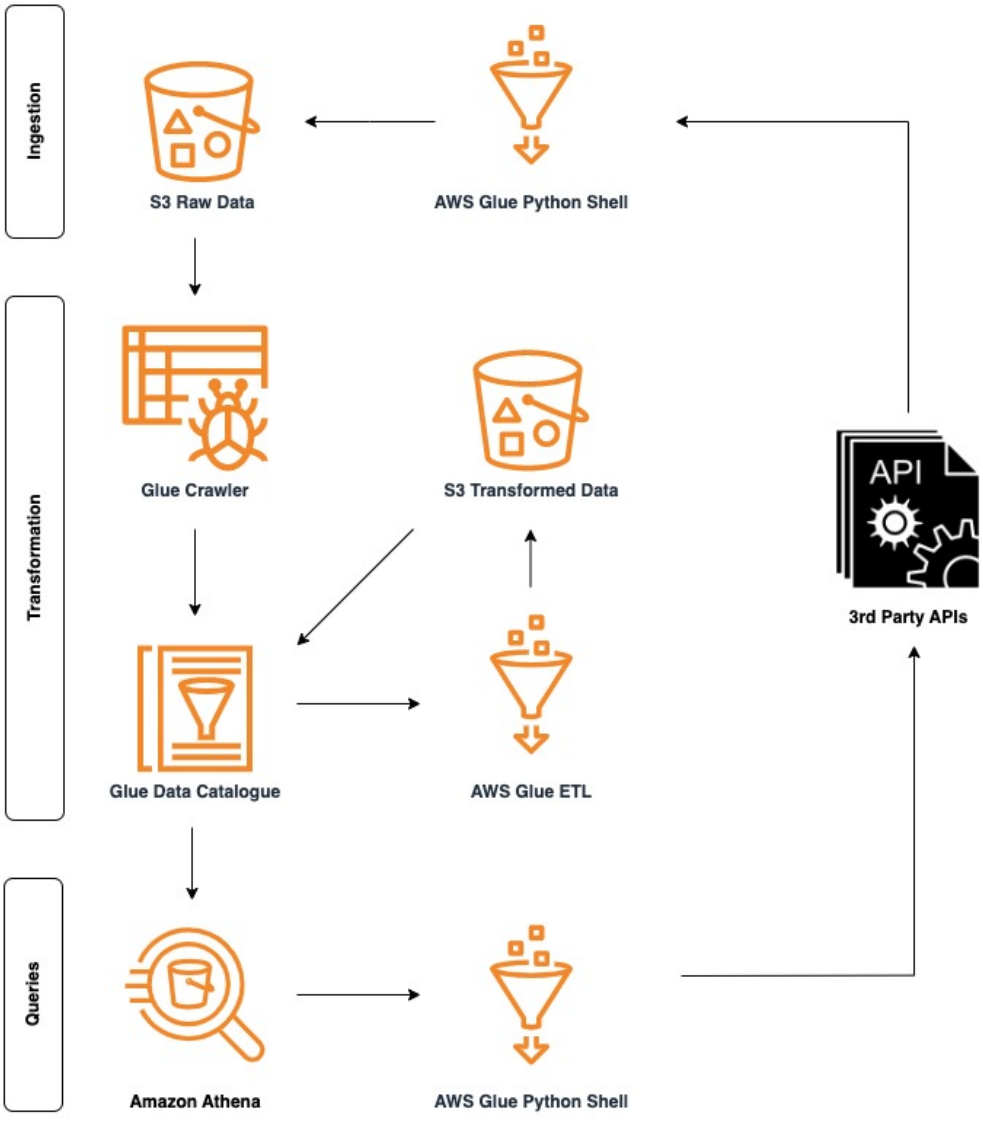


Figure 5.2: Data ingestion, transformation and query process

5.2.1 Data ingestion with AWS Glue Python Shell

AWS Glue is a serverless data integration service that offers both a visual and a code based interface to manage various types of workloads. It can be used to discover, prepare, move and integrate data from multiple sources for purposes such as data analysis and machine learning. Simple Python shell scripts can be scheduled to run with AWS Glue Python Shell and more complex data pipelines can be built and managed with AWS Glue ETL (extract, transform and load). In AWS Glue, customers are charged an hourly fee based on the Data Processing Units (DPU) they use. Apache Spark and Spark Streaming jobs require a minimum of 2 DPU to run, while Python shell jobs can be allocated either 1 DPU or 0.0625 DPU. The price of one DPU at the time of writing is $0.44 per hour with a minimum billing of 1 minute.

Data is ingested into the data lake via scheduled jobs run by AWS Glue Python Shell. These scripts request data from source systems via APIs and write the fetched data into the raw data zone of the data lake. Data could also be uploaded into this section of the data lake manually or by automated file transfers in the case of source systems that do not offer an API. Each source system has its own section inside the raw data zone and data inside these sections is organised in a way that supports efficient querying. For instance, web shop sales data is written into a sales folder under the section of Mycashflow inside the raw data zone. This type of data is frequently queried for given time frames and the individual sales data objects are thus organised into folders by date, broken down by year and month, resulting in a hierarchical directory structure visible in this example path of a sales transaction: S3://raw-data-zone/mycashflow/sales/year=2023/month=01/order=6603/57309.json. AWS Glue offers enhanced support for working with datasets that are organised according to this type of Hive-styled partitions as we will notice in the next sections.

The data is mostly stored as-it-is in the raw data zone with the exception of

minor cleaning done to avoid problems in the consecutive steps of the process. Some of the source systems showed inconsistencies with regard to data types of id fields that were sometimes represented as integer and at other times as string values in JSON formatted data. Similarly, string representations of numerical values, such as product price and weight data, were given consistent decimal formats at this stage of the processing. The ingestion scripts were also used to avoid bringing in unnecessary sensitive data into the data lake, such as personal information of private consumers that could have proven difficult to manage in a compliant way under the General Data Protection Regulation (GDPR).

5.2.2 Data transformation with AWS Glue ETL

The next step in the ingestion and transformation process is to identify schema changes and new partitions for the imported raw data and to create or update AWS Glue Data Catalogues accordingly. *AWS Glue Data Catalogue* acts as a central metadata repository for the Glue environment and enables discovering data directly from S3. Catalogued data is immediately available for search and query from other AWS services such as the interactive Amazon Athena query engine. The data catalogue contains table and job definitions as well as schemas and other relevant information about the Glue environment. AWS Glue crawlers can be used to automatically determine the schema of the data in a given source or target data store and to create or update the metadata in the data catalogue accordingly.

In our data lake solution, we scheduled AWS Glue crawlers to periodically connect to the data sources in our raw data zone. The crawlers were configured to run built-in JSON and CSV classifiers on sample datasets to infer the schema and write the metadata to the Data Catalogue. Once the metadata in the AWS Glue Data Catalogue has been updated, the raw data can be transformed via a Glue ETL job into an optimal format for querying with Amazon Athena. The main objectives of

the ETL job is to modify JSON supported data types to native SQL data types and to re-partition individual objects into suitable compressed subsets of data to reduce the amount of costly I/O operations when querying data from the data lake. String represented timestamps and decimal values are mapped into native SQL timestamps and decimal values. Source system related fields are left without a mapping in order to limit the amount of data in the transformed data zone to only data that is relevant for reporting and analysis.

AWS Glue ETL jobs import data from S3 into so called DynamicFrames and the amount of data written can be limited by using push down predicates based on partitioned columns. This way new raw data can be transformed without reading and filtering all the available raw data in the data lake. Our AWS Glue ETL scripts update AWS Glue Data Catalogue and write the transformed data into the correct S3 bucket in optimal subsets according to given partitioning rules in order to ensure effective queries in the future. Sales data, for instance, is still partitioned by date (year and month), but all monthly sales data of a given year is written into a maximum of five parquet files, resulting in the following directory structure for the first batch: S3://transformed-zone/mycashflow/sales/year=2023/month=01/run-1676253958243-part-block-0-0-r-00000-snappy.parquet.

5.2.3 Data visualisation with Amazon Athena and Power BI

Amazon Athena is an interactive query service that can be used to query Amazon S3 using standard SQL. Amazon Athena is serverless, meaning users only pay for the queries they run and do not need to manage any servers or data warehouses. They can simply point to their S3 storage, provide a schema and run queries without complex ETL processes. It can, however, be used alongside with AWS Glue, in order to automate schema creation with Glue crawlers and to optimise query performance by transforming data with Glue ETL jobs. Amazon Athena supports multiple data

formats such as CSV, JSON, ORC, Avro, and Parquet. Choosing the right data format can affect the cost and speed of queries run with Amazon Athena. At the time of writing, the charge for scanning a terabyte of S3 data for Athena queries is $5, but this cost can be cut up to 90% by compressing, partitioning, and converting the data into columnar formats like parquet.

Amazon Athena is used in our data lake solution to create and periodically update customer and sales related datasets. The predefined queries are run from a scheduled script running in AWS Glue Python Shell. The query results are transformed into meaningful KPIs using pandas - a python data analysis library - and written back into the data lake to act as datasets for business intelligence reporting and system integrations. The effects of partitioning and compression on query performance in Amazon Athena are visible in Table 5.1.

Table 5.1: Amazon Athena query performance experiments

Input format	Partitions	Run time	Data scanned
JSON	no partitions	21.364 sec	105.12 MB
JSON	year and month	6.785 sec	7.80 MB
parquet	no partitions	4.057 sec	1.16 MB
parquet	year and month	4.405 sec	117.84 KB

These experiments were done with an SQL query that joins customer and sales data to retrieve aggregated sales data for B2B customers for a three month period as well as comparison data for the same time period the year before. The amount of data scanned is significantly lower when using partitioning compared to reading all the available data from the data lake. Using re-partitions with a compressed file format will further reduce the amount of data scanned. Even if the cost effects are not significant with this relatively small dataset, partitioning and data compression become increasingly important when the amount of data grows.

Microsoft Power BI is a business intelligence tool by Microsoft designed to bridge

the gap between data and decision-making. It allows users to visualise their data and to share their insights. Power BI can be used to build BI reports and interactive dashboards that integrate vast amounts of data from multiple sources and formats. Power BI integrates well with commonly used enterprise tools such as Excel. Power BI Desktop can be downloaded free of charge and used to create BI reports locally. Source data can be imported from Amazon Athena into Power BI desktop via a special connector. Reports and dashboards can be made available online with Power BI Service. Published content can be kept up to date with Amazon Athena by Power BI Gateway.

We used Amazon Athena connector for Microsoft Power BI to give the case company management access to the data inside the data lake. Once the connections were configured and data models defined, Power BI became a self-service reporting and analysis tool that empowered the management of the case company to ask more complicated questions from their data and to share their insights. In addition to introducing a reporting and analysis tool directly connected to the data lake, we also used the tools provided by AWS Glue and Amazon Athena to create integrations between source systems. Scheduled AWS Glue Python Shell scripts run daily to create and update customer data directly from the MyCashflow web shop into the Monday.com CRM system. Similarly, the CRM customer data is enriched on a weekly basis with sales KPIs for the past year along with a year-over-year comparison per customer. Id mappings are stored back to the data lake in order to create a reliable link between the same entities in these two distributed systems. Figure 5.3 demonstrates the relationship between the sources of data, the data lake and the interfaces used to visualise the data.

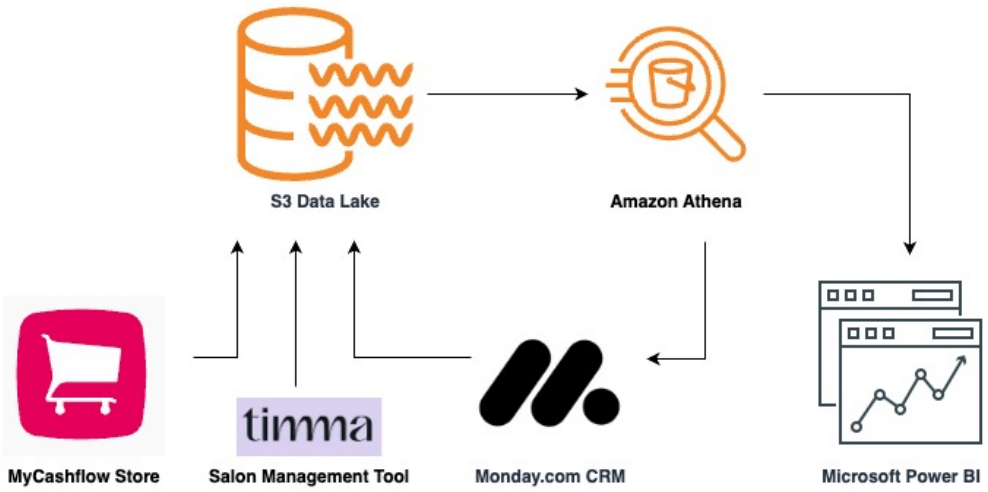


Figure 5.3: Overview of the data lake solution

5.2.4 Identity and access management

AWS Identity and access management (IAM) is used to securely manage identities and access to AWS resources and services. IAM roles can be used to grant people and systems access to AWS resources via fine-grained permission policies. Effective use of these IAM features will ensure that the right people have access to the right resources under the right conditions.

We used AWS IAM to create an admin user for handling administrative tasks and the necessary number of users for the people working at the case company. We also created a service role to be used by AWS Glue when accessing other AWS services. We created a group that was granted access to all the necessary resources for reading the data lake content from Power BI via permission policies. The company users were given access to the data lake by adding them to this group. This way it would be easy to manage access for all the users as a group and easily remove users from the group that no longer needed access.

AWS Secrets Manager is used to securely store credentials, API keys and tokens. This can be useful when trying to avoid hard coding secrets in plain text to applications that need access to other systems. AWS Secrets Manager enables automating

secrets rotation and monitoring secrets usage. We used AWS Secret Manager for safely storing source and target system API keys used by AWS Glue Python Shell while fetching new raw data and updating processed data via systems integrations.

5.3 Validation of the solution

In this study we attempted to identify some of the challenges of data-driven decision-making in small businesses, especially related to data visibility. We built a data lake on the AWS cloud computing platform in order to explore how these challenges could be solved by leveraging cloud computing. In the beginning of the design process the management of the case company were aware of the need to have integrated and aggregated customer and sales data from various systems to support their decision-making, especially in terms of assessing the effectiveness of marketing activities, allocating scarce resources and identifying new business opportunities. The data lake solution was intended to function as a low cost central repository for company data from various distributed systems with automated data ingestion and transformation processes that would remove the need for manual data gathering and processing. The solution was expected to enable easy data modelling and effective data queries, and to provide interfaces for BI reporting, dashboards and analysis.

During this study, we were able to set up the central repository and to create automatic data transfers from two source systems and enable manual data transfer from one source system into the data lake. We were able to automate data transformation processes and schedule queries to create timely aggregate datasets to be used for reporting and integrations. The management of the case company stresses that having proper visibility to data and actively utilising this data as well as communicating about it are essential for accumulating expertise inside the organisation and building trust in customer relationships. They acknowledge that having the cloud based data lake significantly improves their possibilities for taking advantage

of data-driven decision-making, although the actual achieved benefits are still below the full potential offered by the solution.

One of the most visible benefits of the data lake solution was that it enabled displaying and automatically updating refined, timely and reliable customer data and sales KPIs from the data lake inside the tools used by the marketing team in their daily work, without costly and error prone manual data transfers and calculations. Reliability of the integrated customer data and sales KPIs is deemed high. The KPIs are reviewed on a weekly basis to maintain a sense of how things are progressing, which customer relationships need to be focused on and to how to respond to changes. Sales KPIs are also used in support of status discussions with resellers. This used to require manual data retrieval and calculations but the same data is now readily available at all times reducing the time needed for preparation.

The benefits provided by the data lake in terms of assessing the effectiveness of marketing activities were more limited. This type of analysis requires large amounts of data about different types of marketing activities, such as campaigns, sales meetings and training sessions. This data had not been recorded in a systematic way, which made it difficult to import this data into the data lake or create data models that would connect it to sales transactions or other measurable outcomes. The management of the case company realised during the design process that increased knowledge also increases the awareness of the challenges related to data-driven decision-making. They had to establish work flows that would enable accumulating the necessary data for analysis in the future.

Similarly, the benefits of the self service BI reporting and analysis tool, Microsoft Power BI, did not fully materialise during the time span of this study. On the one hand, it provides the management of the case company with flexible access to their data. They can define new data models and create elegant visualisations, but on the other hand, this requires data analysis skills that need to be acquired on the side of

other daily tasks. The potential of this tool was nevertheless deemed high although becoming a proficient user and realising these benefits would take time.

All in all, the management of the case company verified that the data lake solution had had a very positive effect on the daily operations of the company. The costs of running the data lake were considered acceptable and the potential for creating business value through more data-driven decision-making was assessed much higher because of the data lake compared to the situation before it.

6 Conclusions

In this thesis we studied how to enable data-driven decision-making in a Finnish SME by leveraging cloud computing. The main research question was broken into three sub-questions: (1) what are some of the technical challenges hindering data-driven decision-making in small businesses?, (2) how can a cloud based data lake be utilised to solve these challenges? and (3) what kind of an effect does the implementation of a cloud based data lake have on the ability of the case company to take advantage of data-driven decision-making? The first question was addressed in Chapter 3 while the second and third question were discussed in Chapter 5. We conclude with some practical and theoretical implications that emerged throughout the research process.

6.1 Practical implications

In the case study of this thesis we designed a cloud based data lake to enable data-driven decision-making in a Finnish SME. The purpose of the data lake was to address the challenges of data availability by acting as a unified storage solution for company data from distributed systems. We found that a cloud based data lake can be a flexible, low cost solution for an SME, especially for growth oriented sales organisations, that are on the outlook for new business opportunities. The data lake was selected over alternative, more traditional decision support systems, such as the data warehouse. Based on our experience, it can be argued that building a

cloud based data lake is a suitable solution for a small sales organisation, since it does not require heavy ETL related design work in the beginning of the project, but rather allows running small experiments and learning by doing. Cloud based data lakes support simple data ingestion and enable exploring the data and the analytical possibilities it offers without heavy IT investment or predetermined goals.

The availability of cloud platforms makes proprietary tech solutions like data lakes available to small businesses alike. These businesses can opt for the most flexible solution from the start when considering a more traditional decision support system and the modern solutions available on cloud computing platforms. Therefore, the software choices of SMEs may not follow a familiar path seen in large companies that have first invested in data warehouses and then moved on to building data lakes as the requirements for data handling to support data-driven decision-making have changed.

Building a cloud based data lake involves automating data ingestion and transformation routines, which requires data engineering skills, but there is little need for continuous maintenance once the solution is ready. Cloud computing platforms such as the Amazon Web Services, and more specifically its serverless data integration service AWS Glue, can effectively alleviate the challenges of building and maintaining data lakes, especially with its automated schema detection and metadata management. Therefore, with the help of a data engineer, even a small sales organisation can leverage the possibilities of cloud computing.

Our findings suggest that selecting the right software systems can become a critical success factor for SMEs because of the implications it has on the ways of working, both inside the company and with customers, and the financial consequences it bears. Building a data lake can alleviate some of the shortcomings of SaaS business applications used by SMEs, especially in terms of reporting features, since it enables building new data models and creating aggregate datasets based on data from mul-

tiple sources. In this study we addressed these challenges by building a cloud based data lake and found that the solution not only improved the availability of data for decision-making, but also improved the quality of the data. As a consequence, it was possible to have reliable data at hand to support daily decision-making and to respond to changes in the data in a timely manner.

At the same time, we noticed that lack of data analysis skills can substantially hinder the benefits of a data lake solution for an SME. Considerable time investment may go into acquiring the necessary skill set. However, it might be more beneficial in the long run to make these investments instead of pouring money into more sophisticated IT systems and external analytics services that may still not provide the kind of insight necessary to make profound decisions. One could argue that data-driven decision-making is becoming the norm and entrepreneurs as well as small business leaders need to pay special attention to building these capabilities inside their organisations.

Embarking upon a more data-driven decision-making process can influence the ways of working inside an organisation. It may reveal that existing workflows are vague at best and need to become more consistent. The organisation can adopt new workflows that support the accumulation of data that can be used to support decision-making in the future. Therefore, it is important to note that taking full advantage of the possibilities provided by a data lake solution requires developing workflows and organisational culture toward a more data-driven mindset. The benefits of a more data-driven decision-making process enabled by this technical solution are not guaranteed and may materialise during a long time period as a result of various iterations of design work.

6.2 Theoretical implications and limitations

The data lake is a relatively new concept with growing interest among practitioners as well as academia. Several data lake prototypes have been proposed and many relevant research problems have emerged. Some of the remaining challenges relate to metadata management, data quality, data provenance, metadata enrichment, data preparation, dataset organisation, modelling, data integration and related dataset discovery [13]. We contribute to this work by providing a case study based solution designed to meet the needs of a sales oriented SME.

In this study we followed design science principles and built an IT artefact, the cloud based data lake, through an iterative design process in order to solve a real-life business problem faced by our case company. We aimed to solve a relatively common problem hindering data-driven decision-making - the integration and visualisation of distributed company data - with a relatively new technical solution more rarely used in the SME context. It is important to note that while our findings are tentatively encouraging, they are derived from a single case study and may not be generalised beyond the context of the study. Our conclusions are mostly practical, but there are some theoretical implications as well concerning data lake architecture, SaaS adoption and data-driven decision-making in SMEs.

In the literature on data lakes, they are frequently offered as a solution for handling large quantities of fast moving data of various formats, also referred to as big data. Although data lakes are arguably well suited for working with big data and also SMEs are increasingly finding themselves working with large, heterogeneous datasets, the benefits of a data lake over a more traditional data warehouse solution in this case study had to do with other aspects than the volume of data. For an SME, a cloud based data lake can be a low cost unified storage system that is closer to an integration platform than a big data repository used for data discovery and analysis.

Data lake literature offers different types of architectural models for data lakes. Our data lake architecture is an example of a functional data lake architecture with different zones for raw data, transformed data and query results. We found that while data lake architecture matters also for SMEs, the optimal amount of zones needed to keep the data organised may be more limited when there are fewer data sources and a restricted amount of data. We have, for instance, only one zone for raw data instead of a separate transient loading and raw data zones. Similarly, we have only one transformed data zone instead of having both, a refined data zone and a trusted data zone.

Data lake governance, and metadata management in particular, emerge as important themes in data lake literature. In our cloud based data lake, we do not have a separate governance zone or complicated metadata management processes. This could be considered a limitation of our solution, but since it has not produced any problems, one might also argue that data lake governance and metadata management are less critical when building data lakes for SMEs with limited amounts of data. Furthermore, cloud computing platforms offer automated tools for data governance and metadata management.

During this study we discussed the software choices made by the case company with special emphasis on the benefits and limitations of the selected systems. We found that the SaaS model was considered attractive because it provided sufficient functionality without costly investment in infrastructure and maintenance activities. The reliability and cost savings provided by the SaaS model were seen as more beneficial than the added adaptability or sophisticated features of customised solutions. Besides sufficient functionality, the selected software systems needed to support flexible team working irrespective of time and place. These findings are similar to those suggested in the literature, although the role of cost reductions is emphasised.

Literature on maturity models suggest using assessment tools to evaluate the

state of data-driven decision-making in SMEs. We could have used this approach to assess more rigorously the effects of the data lake solution on the ability of the case company to take advantage of the possibilities of data-driven decision-making. Our findings suggest that the implementation of a data lake can help a small business to progress from such lower levels as uninitiated and awareness to higher levels such as proactive adoption and integral embracement, especially on the dimensions of data availability, data quality and information use. We were, however, not able to perform a full state assessment of data-driven decision-making in the case company before and after the data lake implementation due to the limited resources and time scope of the study.

References

[1] "How amazon makes money", Investopedia, Ed., 2022.

[2] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research", *MIS Quarterly*, 2004.

[3] A. Malviya and M. Malmgren, *Big Data for Managers: Creating value*. Routledge, 2019.

[4] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making", 2013.

[5] R. Vidgen, S. Shaw, and D. B. Grant, "Management challenges in creating value from business analytics", *European Journal of Operational Research*, 2017.

[6] "SBA Fact Sheet - Finland", European Commission, Ed., 2019.

[7] X. Parra and X. Tort-Martorell, "Assessment of information-driven decision-making in the sme", *International Conference on Information Quality (ICIQ)*, 2016.

[8] X. Parra, X. Tort-Martorell, C. Ruiz-Viñals, and F. Álvarez-Gómez, "A maturity model for the information-driven sme", *Journal of Industrial Engineering and Management*, 2019.

[9] N. Miloslavskaya and A. Tolstoy, "Big data, fast data and data lake concepts", *7th Annual International Conference on Biologically Inspired Cognitive Architectures*, 2016.

[10] C. K. Leung, "Data science for big data applications and services, in: Big data analyses, services, and smart data, aisc 899", 2021.

[11] D. Taniar and W. Rahayu, "Data lake architecture", *The 9th International Conference on Emerging Internet, Data and Web Technologies (EIDWT)*, 2021.

[12] "Why are data lakes the future of big data?", Oracle, Ed., 2019.

[13] R. Hai, C. Quix, and M. Jarke, "Data lake concept and systems: A survey", 2021.

[14] E. Malinowski and E. Zimányi, *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications*. Springer, 2009.

[15] M. Jürgens, *Index Structures for Data Warehouses*. Springer, 2002.

[16] "Cost modeling data lakes for beginners: How to start your journey into data analytics", Amazon Web Services, Ed., 2020.

[17] F. Nargesian, E. Zhu, R. J. Miller, Q. P. Ken, and P. C. Arocena, "Data lake management: Challenges and opportunities", *Proceedings of the VLDB Endowment*, 2019.

[18] D. E. O'Leary, "Embedding ai and crowdsourcing in the big data lake", 2014.

[19] C. Giebler, C. Gröger, E. Hoos, H. Schwarz, and B. Mitschang, "Leveraging the data lake: Current state and challenges", *Proceedings of the 21st international conference on big data analytics and knowledge discovery*, 2019.

[20] F. Ravat and Y. Zhao, "Data lakes: Trends and perspectives", *30th International Conference on Database and Expert Systems Applications (DEXA)*, 2019.

[21] P. Sawadogo and J. Darmont, "On data lake architectures and metadata management", *Journal of Intelligent Information Systems*, 2020.

[22] "Angling for insight in today's data lake", Aberdeen Research, Ed., 2017.

[23] J. Dixon, "Pentaho, hadoop, and data lakes", 2010.

[24] Y. Zhao, "Metadata management for data lake governance", Ph.D. dissertation, Univertité Toulouse I Capitole, Institut de Recherche en Informatique de Toulouse, 2021.

[25] P. Gupta, A. Seetharaman, and J. R. Rudolph, "The usage and adoption of cloud computing by small and medium businesses", *International Journal of Information Management*, 2013.

[26] Greenlips Finland Oy, "Tarinamme", 2022. [Online]. Available: `https://greenlipsbeauty.com/page/27/tarinamme`.

[27] MyCashflow, "Tietoa meistä", 2022. [Online]. Available: `https://www.mycashflow.fi`.

[28] TimmaOy, "Enemmän aikaa keskittyä olennaiseen", 2022. [Online]. Available: `https://join.timma.fi`.

[29] Monday.com, "So how did monday.com come to be?", 2022. [Online]. Available: `https://monday.com/p/about`.

[30] Procountor, "Procountor taloushallinto", 2022. [Online]. Available: `https://procountor.fi/procountor`.