Bachelor's thesis

Information and Communications Technology

2020

Phuong Pham

A CASE STUDY IN DEVELOPING AN AUTOMATED ETL SOLUTION

– Concept and Implementation

TURKU AMK
TURKU UNIVERSITY OF APPLIED SCIENCES

BACHELOR'S THESIS | ABSTRACT

TURKU UNIVERSITY OF APPLIED SCIENCES

Information and Communications Technology

2020 | 43

Phuong Pham

A CASE STUDY IN DEVELOPING AN AUTOMATED ETL SOLUTION

- Concept and Implementation

The study focuses on the implementation of an automated extract, transform, and load (ETL) solution for the commissioner company of this thesis, a global spares supply company. The commissioner company has an existing automated solution that still has some manual steps in the progress and limitations in extracting transparent tables in its enterprise resource planning (ERP) system and other online sources. The objective of this thesis was to introduce and implement the ETL process that transforms raw data from multiple data sources into meaningful and valuable information in the data warehouse. The research approach for this thesis is action research. This is a combination of taking action and doing research for the given problem. The definition of ETL was examined and its implementation areas were studied based on the combination of quantitative and qualitative methodology. Data cleaning and the application of data mining techniques were also implemented in the process to extract knowledge.

The thesis was carried out within the scope of a global spares supply company. The study was carried out with qualitative research interviews in the commissioning company, study of the existing process along with analyses on the performance of the existing process. The interviews were used to gather information about the views of the interviewees about ETL and its current challenges.

For the development of the system, this thesis explained a deployment process, introduces libraries, and shows how to utilize these libraries for data integration. The deployment process was built and reviewed by the case company and adjusted to better meet the case company needs. The outcome of this thesis is the automated ETL scripts released on production for the case company.

KEYWORDS:

Extract, Transform, Load, Scripting, Integration System, Data Integration, ETL, Data Warehouse.

ACKNOWLEDGMENTS

I am finishing this thesis during the internship time at KONE Industrial Lt.

At first, I would like to warmly thank my thesis supervisor, Senior Lecturer Tina Ferm for all the guidance and supports throughout this bachelor's thesis project. She has been patient to review and comment on many versions of this thesis. During this journey, I also wish to thank Jani who trusted in my ability to develop this study and allowed me to explore and learn. And I want to thank Nnamdi for all guidance throughout the research process and detailed and useful feedback. The guidance I received was invaluable. My colleagues, Anna and Ben always have time to discuss the thesis. Thank you all for always supporting me during this journey. Finally, I would not be where I am today without the unconditional support and love of my mother and Thuan, who have done a lot of things for me and have always supported me in pursuing my study dreams. Thank you for all of your endless love, ongoing support, and encouragement regardless of physical long-distance.

Finland, April 2020
Phuong Pham

CONTENTS

FIGURES

TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
API	Application Programming Interface
ETL	Extract, Transform and Load
ERP	Enterprise Resource Planning
GSS	Global Spares Supply
GUI	Graphical User Interface
IoT	Internet of Things
HTML	HyperText Markup Language
ML	Machine Learning
NLP	Natural Language Processing
OLAP	Online Analytical Processing
SAP	Enterprise Resource Planning software
PDF	Portable Document Format
RDC	Regional Distribution Center
WWW	World wide web

1 INTRODUCTION

This thesis was carried out as a research project on the topic “A case study in developing an automated ETL solution – Concept and Implementation”. This research, which is based on theory and existing knowledge, is about data integration modeling, its process capability and improvement areas, and benefits to businesses. The work is expected to contribute to the commissioning company’s needs for timely detailed information and a sufficient amount of data gathered for big data analytics development. Figure 1 outlines the focus of the thesis.

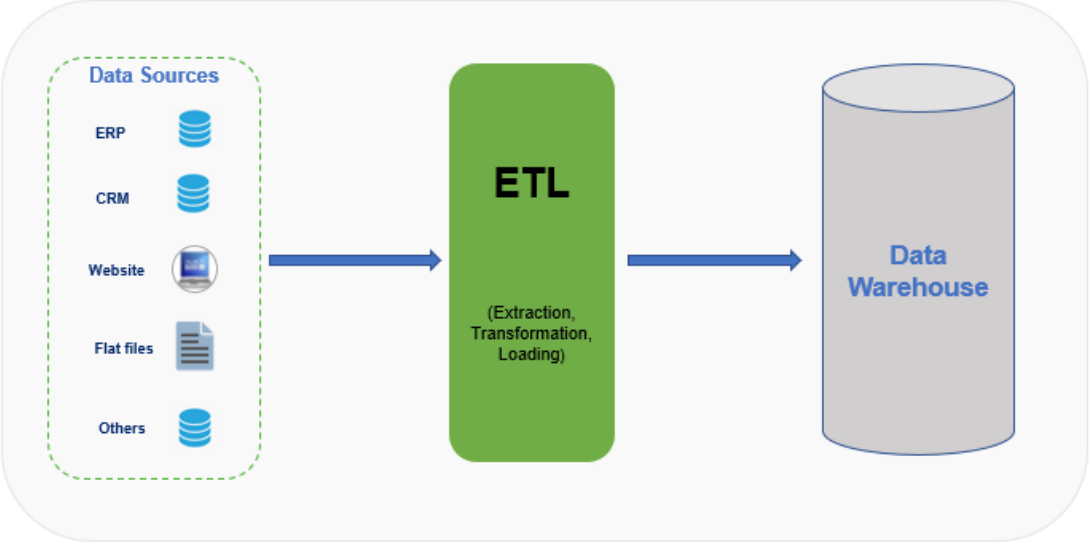

Figure 1. An overview of the ETL process.

This thesis is divided into five chapters. The initial part provides the background to the research including an overview of the data integration process, motivation research aim, research questions, methodological approach, and outline (Chapter 1). The introduction has been divided into multiple subjects to introduce the ETL process in general, which is also called data integration and its solutions for the case company. Data integration is mentioned, in particular, the methods of extracting, transforming, and loading data. Next is the literature review about data source systems, data warehouses, and the integration tool introduced and used in this thesis. Chapter 2 also discusses some of the data mining techniques that are used to leverage data quality hoping that technology-assisted data analysis can provide the diligence and reliable quality control needed for businesses. The next chapter, Chapter 3, overviews the methodology discussing quantitative and

qualitative research being used in this thesis. In the fourth chapter, the implementation of ETL processes was performed with detailed frameworks. The concluding chapter, Chapter 5, discusses the contributions of applying new developments to enhance the performance of ETL processes in the case company and limitations of the research.

1.1 Background

The Extract-Transform-Load (ETL), is defined as a process, not physical implementation. This is the core function of data integration and warehousing operations. Managing the data transferring between systems is a major challenge and a serious concern for every organization. The complexity of an ETL process is mostly depending on the size and characteristics of each organization, the cost, and the location of storage. Properly and efficient ETL systems can bring huge benefits to businesses and be used for crucial business decisions. This tool is designed to meets high demands for extracting and storing a huge volume of data generated every minute with high quality and consistency standards. In order to achieve that, ETL performs by collecting data from various source systems, integration, and delivering source data into a destination system or target platform. This platform can be called Data Warehouse (DW). Besides, one of the top ETL tools expands its functionality for data cleansing, processing, customization, and reformatting. The whole of the ETL process includes three steps, namely extract, transform, and load.
The ETL process can be completed by manually editing the data or having automated processes through automation scripts or programs.

1.2 Motivation for the research

Nowadays, the demand for sufficient data is exploding and will continue developing further in all industries. In other words, data has a huge impact on the way businesses operate by changing insights into marketing, customer behavior, prospects, and competitors through technological evolution. Indeed, data is at the heart of the most influential concepts in business management and a key factor in business success. For instance, by collecting enormous amounts of transactional data from multiple sources, mainly enterprise resource planning (ERP) system, the business can track customers' purchase history as well as assess operational situations globally. However, storing data

is not enough for the organization, but it is crucial that the data is integrated into a single place. To achieve that, the organization must carry out an ETL process.

Manual ETL processes cause a great amount of manual work through double inputs into the systems using external tools such as Excel. In addition, as data can be collected from many different sources and increases to several billions of rows, a classic ETL tool can work but might lead to troubles in more complex scenarios. Thus, there is another concern that companies could not meet their needs for timely detailed information for making-decision with a manual ETL pipeline. Hence, automated data integration helps to avoid misinformation, possible human errors, gain overall visibility, and reduce manual work for the long term. The majority of global companies spend considerable resources on building the ETL system and processes. The ETL system needs to demonstrate the ability to integrate data from various sources with different types of data and then, load and store successfully in a target data warehouse.

The main aim of this implementation is built upon the enhancement of current ETL processes to support some missing functions in the case company and increase the performance of ETL processes to full capabilities in extracting data from multiple sources.

1.3 Research Questions

This research study aims would solve the problems in the case company by answering the questions as stated below:

How does the existing ETL processes in the case company look like?

What kind of current issues and development areas are there?

How can the new improvements benefit the case company?

1.4 Research methods

The workflow of this study is divided into four main phases. The initial phase is the knowledge gathering about ETL, data sources, data warehouses, and data processing techniques. The next phase is to carry out interviews with the team members who are working closely with ETL projects and refine the results after interviewing. Then, from the

analysis of the results of the interviews, the solutions are developed and implemented to the problems. The final phase examines the outcome and delivers a conclusion.

1.5 Focus and limitations

This research study focuses on introducing ETL processes and developing additional functions for the current ETL tools for the case company. This thesis does not discuss the scope of the whole business. This implementation is only carried out on the scope of the Global Spares Supply (GSS) unit in the case company in Finland.

2 LITERATURE REVIEW

This chapter explains technical terminologies and fundamental technologies related to the ETL process including data source system, structured and unstructured data, data warehouse, data processing techniques that have applied to this study, and how they are combined to implement.

2.1 ETL process

This chapter will expose the extraction, transformation, and load (ETL) process. This definition can be understood as both a simple and complicated subject. In its easy form, the term “ETL” is initially seen as the actions of getting data out of the source and load it into the data warehouse. In fact, ETL is a complex combination of process and technology that consumes a significant portion of the large data warehouse enlargement efforts and needs the skills of trade analysts, database (DB) designers, and application developers. (Gajare & Rangdale 2017). In particular, these tools are typically pieces of software responsible for the extraction of raw data from multiple sources, performing transformations it into a format that can serve business needs, and loading it into a data warehouse. This is a crucial stage in the data preparation process and it supports providing clean, reliable, consistent, and accurate data in an analytical database. Data mining techniques are sometimes included in this ETL process to turn the data into knowledge if needed.

Basically, the ETL process contains three discrete functional fundamentals:

- Extract: the main objective of this step is to retrieve all the required data from multiple sources. Therefore, the first step determines which source systems, the velocity of each source, and the priorities of data flow.
- Transform: After data extraction, transformation brings clarity and clean data based on a set of rules from basic to complex transformation. This step also involves the following tasks such as cleaning, filtering, splitting, and gathering data using aggregate and summarization function, generates new calculated values, and applies advance validation rules. Transforming selected data into forms appropriate is also considered as the most complicated part of integrating data.

- Load: the last phase of the process, targets source, and refresh rates also need to be determined before execution. It is vital to ensure that the load is performed correctly into the target database or storage.

The important functions such as parsing, data enrichment, setting velocity, validation figures need to be understood and considered carefully and placed in the context of the specific needs of a company.

Loading data is a process of bringing data into the system from any external system using an external file or by establishing a connection to a live data source.

- Manual ETL process

Without ETL tools, the process of extraction, transformation, and loading must be done manually by those who are working in the data warehouse department. Collecting data from multiple sources and joining all the tables are highly time-consuming and might cause common data quality issues and data security. All these works need to be done before any analysis could even be performed. Another concern is that the firm could not meet end-users' needs for timely detailed information for making-decision, specifically corporations having thousands of thousand transactions per day.

- Automated ETL process

As the amount of received data will easily grow to very huge amounts, this could not be controlled and analyzed by human beings. An automated ETL tool typically provides a time-saver tool for data extraction and preparation for the next steps in analytics and visualization. This helps to enhance the quality of target data by detecting errors and inconsistencies from data and remove it. An automated ETL process is operated through scripts. The script is a computer program has a sequence of commands and functions to process data. It can be scheduled and triggered to run on the servers as desired. Most ETL operations run overnight, and data is ready as end-users arrive for work in the morning. ETL process can be run periodically such as monthly, weekly, daily, etc.

In general, almost all of the ETL processes encompass four phases: raw layer, staging layer, schema layer, and aggregating layer.

- Raw layer: the raw data is coming from the source system is written and copied to the staging area without adjustment or enrichment. Data extracted from

structured source systems usually are written to flat files such as TSV, CSV, XLSX, XML, etc.

- Staging layer: the raw data from the raw layer will not load directly into the target data warehouse. Raw data first is transformed in this step and all transformations are stored in staging tables.
- Schema layer: all the data is cleansed, preprocessed, enriched, and transformed to its final form in this step. Then, these are stored in destination tables which are ready to load into the target data warehouse.
- Aggregating layer: is to aggregate data from the full dataset and optimize data for analysis. This can improve report performance and is allowed to add business logic to calculated measures.

One of the most important aspects when undertaking an ETL process is the selection of the ETL tool. Almost large-scale companies choose ETL packages are offered by the world's technical suppliers such as Oracle Data Integrator, IBM Infosphere DataStage, Microsoft SQL Server Integration Services – SSIS, etc. On the other hand, there are other ways to develop ETL tools, companies can build their own tools depending on their demands and needs to have greater flexibility for customization. The common languages to use for this programming is .Net, Java, Python.

2.2 Data source system

A data source, in the context of computer science and computer applications, is the location where data that is being used comes from. (Technopedia 2020). Usually, the data source for computer programs could be flat files such as excel, spreadsheets, XML files or websites, software applications, etc. Data is kept in structured and unstructured formats. Integrating structured and unstructured data involves tying together common information between them, which is probably represented as master data or keys in structured data in databases and as metadata tags or embedded content in unstructured data. (Reeve 2013). Structured data is essentially stored in enterprise systems (ES) including enterprise resource planning (ERP) and customer relationship management (CRM) to record all transactions that have been undertaken.

One of the source systems that are desired to extract data is production operational systems and ERP, in particular. However, to avoid negative impacts on the production performance of production operational systems, the ETL process is not set up directly or

a component on any ERP system. Enterprise Resource Planning (ERP) is the integrated management of main business processes, often in real-time and mediated by software and technology. (Wikipedia 2020). ERP is usually referred to as a category of business management software typically a suite of integrated applications that an organization can use to collect, store, manage, and interpret data from many business activities. (Wikipedia 2020). This system is designed to support the management and operations of an organization. Companies store large amounts of data on different platforms, thanks to ETL, the organization can get a global view that allows it to make better strategic business. There are several global vendors of ERP software such as SAP, Oracle Application, Microsoft AX.

Apart from ERP systems, there are several internal management tools are commonly used in corporations e.g Outlook email, Microsoft SharePoint. They are also considered as important information for businesses to keep track of project status, internal communication, and document management tool. SharePoint Online is a cloud-based service that helps organizations share and manage content, knowledge, and applications to empower teamwork, quickly find information, and seamlessly collaborate across the organization (Microsoft 2020). Moreover, it offers plenty of functionalities such as automated workflows consisting of sending emails automatically and updating task status.

Flat files are files having no internal hierarchy, can consist of single table data including rows and columns. They are simple data files in text or binary format. Flat files can be transactions, time-series data with excel format, CSV, TSV, XML, etc. And if the script runs on operation system such UNIX, data can be standardized to standard code as American Standard Code for Information Interchange (ASCII). It is called the ASCII flat file.

Another essential data source for businesses is online sources such as the content of the webpage, social media sites, and online publications. Scraping data from these online sources will be greatly beneficial for business owners. This can provide a greater understanding of customers, competition, and the way customers react in the market as well as help businesses predict market trends. This way keeps businesses up to date in a quick way about customer behavior and being more competitive in the market. In simple terms, web scraping is the technological art of mining needed information on websites to collect data. There are two main approaches to automated data retrieval in

real-time within minutes are either to application programming interface (API) service or to utilize open-source libraries that already exist for HTML scraping.

An application programming interface (API) allows programmers and developers to both deliver content to and receive content from the web's larger, more highly used services. (Michel 2013). Developers can use API to build tools and scrape data to create valuable, user-friendly services. Usually, large websites or e-commerce websites offer API keys for developers who wish to fetch their webpage such as Twitter API, Yelp API, Flickr API, etc.

However, some websites do not publish API or it is limited in some ways, there is still an alternative way to API calls for data retrieval is HTML scraping. Web scraping provides a streamlined way to collect data from web pages. Web sites are written using HTML, which means that each web page is a structured document. (Kenneth 2020). The website HTML code is parsed into a programming language using open-source libraries such as R, Python, C#.

Multimedia data is data to be scraped from online social media e.g. tweets, posts, captions. They are usually considered as complex data with high dimensionality and inconsistent structure. Therefore, they are required to advance data processing and mining techniques to unhidden patterns hidden in their data and extract knowledge.

2.3 Data warehouse

Data warehouse (DWH) is a large store of data collecting from a wide range of sources for analysis and reporting and an important asset for any organization. Data warehousing architectures are designed to have consistent data available for the entire organization to use for analysis, to format data particularly for analysis and reporting purposes, to take the stress of analytical reporting needs off the operational systems, and to allow for historical snapshots data. (Inmon 1992). In other words, “Data warehousing” is a practice in data management whereby data is copied from various operational systems into a persistent data store in a consistent format to be used for analysis and reporting. (Revee 2013). Indeed, it is a relational database that is designed for analytics consists of online analytical processing (OLAP). A data warehouse is traditionally located on-premise, is also called an in-house server. Nowadays, cloud-build technologies and hybrid cloud data warehouse are overwhelming by its efficiency and security. In general, data warehouse consists of data sources from the operational and transactional system as ERP, CRM, other enterprises systems, and online external sources. Figure 2 presents the set of layers that data warehouse work.

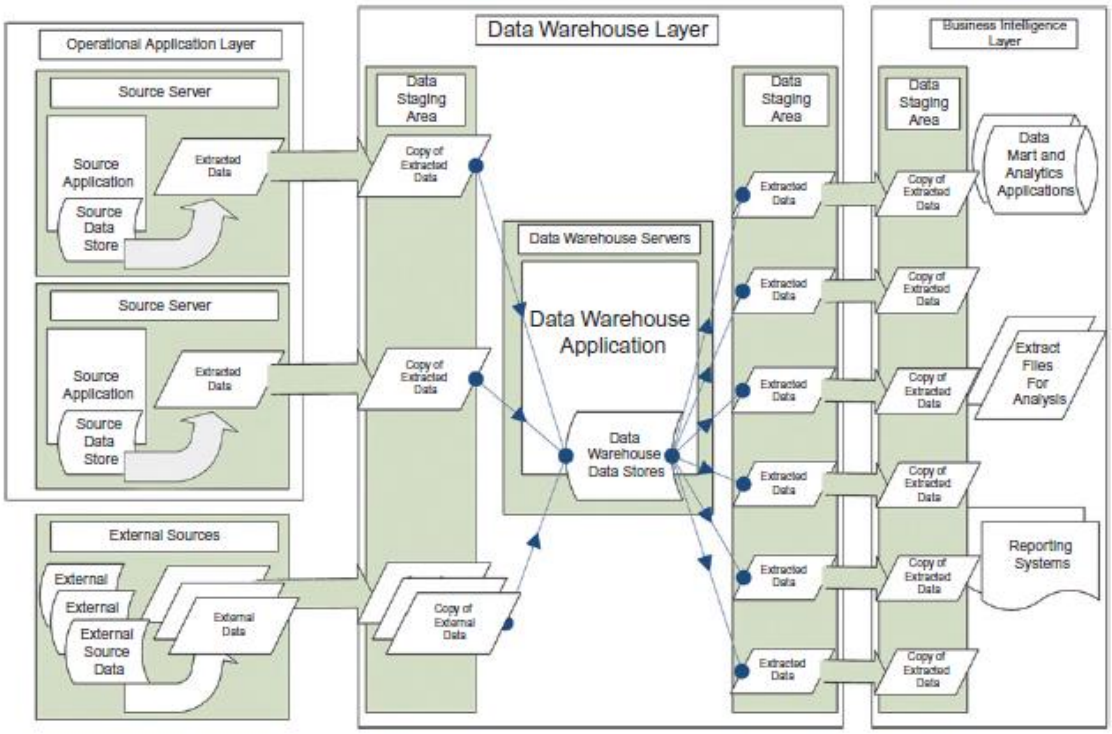

Figure 2. Data Warehouse Data Flow (Revee 2013).

Through the ETL process, data passes from multiple source systems (operational application layer) to data warehouse and then to the business intelligence (BI) layer. Another important thing to an implementation data warehouse is the classification of the types of input data.

The diversity of data types to be transferred into data warehouse are treated differently. And the time intervals of the loading cycle are also conducted on a regular basis, which also depends on the types of data.

2.4 Data preprocessing and mining overview

Data preprocessing is vitally needed to get meaningful information from huge, inconsistent, and noisy set of data. Specifically, raw data is extracted from multiple sources with diverse formats, the quality of the data is insufficient and inconsistent, which seems almost inevitable consequence. Data preprocessing is an important issue for both data warehousing and data mining, as real-world data tend to be dirty, incomplete, and inconsistent. (Han & Kamber & Pei 2006). It can help to improve the quality of data and this is important to every organization for many reasons. Figure 3 shows the steps of data processing.

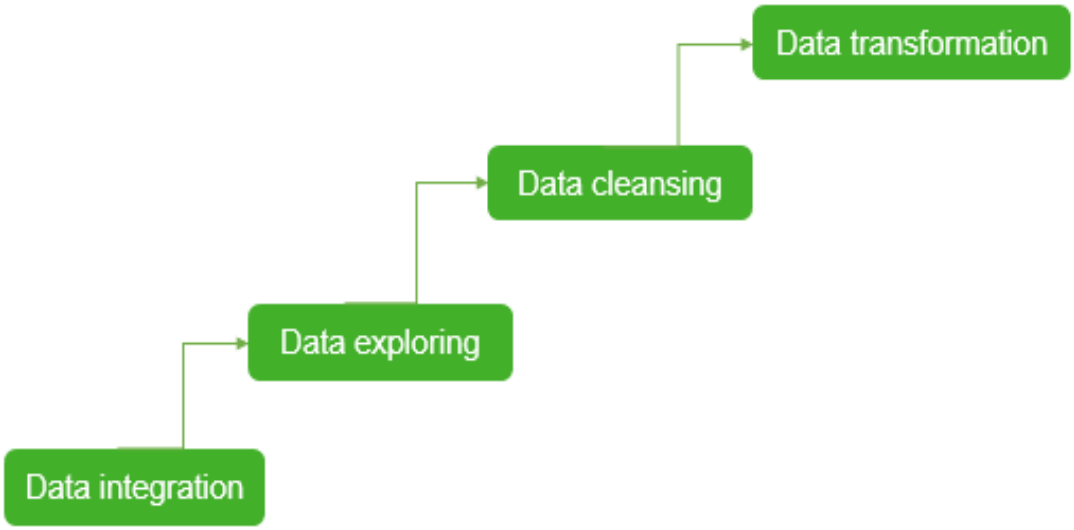

Figure 3. Data preprocessing flow.

Data preprocessing techniques include data integration, data exploring, data cleansing, and data transformation. More importantly, this technique is crucial for unstructured data that is practically useless for analysis of any kind without preprocessing and fitting into particular categories.

After loading data from multiple sources, data is stored and ready for further data preprocessing techniques. Then, exploration generates information that is used to better understand and uncover initial patterns and characteristics of the set of data. Data exploration can be conducted with a combination of manual methods and automated tools for both unstructured and structured data.

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data (Wikipedia 2020). The workflow of this stage is a sequence of three main steps including inspecting unexpected, missing and noisy data, cleaning detected abnormalities and correctness inaccurate, and inconsistent data.

Data transformation is performed to convert and consolidate cleaned data to a uniform layer or forms appropriate as the required format of a destination system. However, due to the complexity of data types, data mining techniques might be involved in data processing to ensure the quality of data before the transformation stage.

Data mining refers to extracting or “mining” knowledge from a large amount of data. (Han & Kamber & Pei 2006). In other words, data mining is the process of discovering patterns and relationships in large volumes of data by using methods from the areas of computer science, statistics, and artificial intelligence. (Clifton 2010). Data mining terminology first was introduced in the 1900s, but it is only becoming a trendy term in recent years in the fourth industrial revolution. Data mining functionalities are varied from descriptive mining tasks to prediction mining tasks. The descriptive technique is used to obtain insights and useful information from raw data, while the main objective of predictive mining is attempting to do predictions for target value based on available data. Depending on the kinds of data and business requirements, the different techniques are chosen to apply. Lately, a new series of data mining techniques have been developed and implemented as deep learning. Those techniques can be mined in text data, but also images and sound data. And specifically, Natural Language Processing (NLP) is a powerful method

that allows machines to create and analyze the human language. NLP is being used for various applications such as text classification, index and searches large texts, automatic machine translation, speech understanding, information extraction, knowledge acquisition, and text generations.

3 METHODOLOGY

This chapter overviews the methodology discussing the single case approach, selection of ETL framework, and the implementation being used with the case company.

3.1 Single case study

A single case is used to provide a practice insight into the application of the ETL process for a global spares supply company. According to Gibbert (2010), single case studies are considered as a greatly valuable approach in the area of applied research. And Rowley (2002) states that a case study is an empirical inquiry and that it is investigating a contemporary phenomenon. Indeed, a case study can reflect the reality of the matter.

According to Yin (2014), a case study can be qualitative, quantitative, or a combination of both of them. In other words, qualitative and quantitative research methods can succeed one another or they can be used simultaneously to compensate for the weaknesses of each other (Cooper & Schidler 2012). Quantitative data collection methods mainly focus on observation and data analysis while qualitative data collection methods are about interviews, case studies along with literature review and theory. The combination of both research methods to be applied in this study provide more value for the outcome of this study. Besides, the research approach for this thesis is action research. Action research differs from that approach by having practical difficulties or issues which need to be addressed (Blichfeldt & Andersen 2006). Indeed, action research in an organization aims to improve the current process through in-depth research and best practices. In simple words, action research is a combination of taking action and doing research for the given problem.

The case company and its spare parts unit are one of the giant global suppliers around the world. This study provides practical insights into the data integration process of the Global Spare Supply unit. Thus, this research is hoped that the implementation of automated ETL processes can be helpful in broader scopes in other companies in different industries.

3.2 Data collection and analysis

This thesis consists of four main steps as mentioned above in section Research methodology. This section outlines the data collection approaches. After gathering the required knowledge for the study, the interviews were carried out with KONE GSS Business Development team members who are working closely with ETL projects.

3.2.1 Interviews

Face to face interviews are a staple method used in qualitative research. Qualitative method is most helpful in solving the research questions, as it helps provide in-depth answers from a focused small group of people who manage and directly use the system after implementation (Wei & Liou & Lee 2008). Thanks to the qualitative research method, people can get a greater understanding of phenomena by answering detailed questionnaires. It usually focuses on the “why” and “how” rather than “what” questions.

The interviews were performed in the Global Spares Supply (GSS) office in December 2019. This is face-to-face interviews mainly consist of open-ended questions in a conversation format. The same questionnaires are given to all participants. The interviews were carried out in English and documented. The purpose of interviews was to learn more about the existing ETL processes in the GSS unit and find out problems as well as potential improvement areas. The structure of the interview consisted of:

- General introduction of interviewees
- Evaluation of ETL processes from the point of view of each interviewee
- Expected features or functions of ETL processes.
- The benefits of automated ETL processes can bring

3.2.2 Data collection

The data was gathered from individual interviews, observation, and the company’s documentation, which is traditional methods used to carry out qualitative research. Meanwhile, the quantitative collection is also applied by assessing the current state of the existing ETL processes regarding this study, its performance (time, volume, and speed) and the needs for it. The most important sources for the ETL process are ERP

systems, in particular SAP system, consists of structure and transparent tables. Other main sources for business analytics such as websites, flat files, social media, etc were examined to evaluate the performance of the current ETL process.

3.3 Reliability

The reliability of this study is based on systematic data collection and documentation. Data collection applies the combination of qualitative and quantitative methods. There are three important parts regarding data collection: users' feedback regarding the current ETL processes, the figures in performance measurement of the existing ETL process, users' expectations on the desired feature of automated ETL solutions.

The performance evaluation figures were from the testings directly on each ETL process towards every data system source and data warehouse. Also, some figures can be recorded in the log files that consist of messages about the success or failure of an event occurring in the ETL processes. Indeed, this information is assumed to be correct.

Feedbacks regarding the current ETL processes, the interviews were carried out with senior employees who have solid background and years experience working closely with the automated data integration process. Therefore, their opinions were most reliable, important, and practical. Thus, all the output and results were observed and reviewed by supervisors, this will support to increase the reliability of this study.

4 FINDINGS OF THE STUDY

This chapter gives an overview of the case company, its spare parts unit, and the company needs for automated ETL processes. Then it discusses the selection of data integration tools.

4.1 Case company overview

KONE Oy, the Finnish company, was founded in 1910 and is headquartered in Helsinki, Finland. The company is a global supplier in the elevator and escalator industry, employing approximately 60,000 personnel worldwide and in 2019 had annual revenue of 9,982 billion euros (KONE 2019). KONE's sales grew 10.0% as compared to the prior year (KONE 2019), the sales were showing the company to have a healthy rate of growth. The firm provides elevators, escalator, and automatic building doors, as well as solutions for maintenance, and modernization services (KONE 2019). Figure 4 below shows the overview of KONE's presence in the world.

Figure 4: Overview of KONE in 2019 (KONE 2019).

The company has divided its business lines into new equipment business, maintenance and modernization. In the new equipment business, KONE offers innovative and sustainable elevators, escalators, automatic building doors, and integrated access control solutions to deliver the best people flow experience. (KONE 2019) In maintenance, we ensure the safety and availability of the equipment in operation, and in modernization, we offer solutions for aging equipment ranging from the replacement of components to full replacements. (KONE 2019) According to KONE annual review in 2019, new equipment business occupied 53%, maintenance accounted for 32%, and only 15% in modernization. All these figures were shown in below pie chart in Figure 5. It can be seen clearly that the new equipment business had a bigger share than the service equipment business.

Figure 5: KONE's businesses in sales in 2019 (KONE 2019).

With the slogan of 'Dedicated to People Flow', the company commits to deliver innovative solutions and good customer service with the purpose for people to move smoothly, safely, and comfortably and without waiting in buildings in an increasingly urbanizing environment (KONE 2019). The company has also introduced four ways to win with customers:

- Collaborative innovation and new competencies

- Customer-centric solutions and services
- Fast and smart execution
- True service mindset

KONE's strategic targets are to have the most loyal customers, be a great place to work, grow faster than the market, have the best financial development in our industry and be a leader in sustainability. (KONE 2019)

4.1.1 Global Spares Supply unit of the Case Company

The Global Spares Supply (GSS) unit is the focus of the research. The Global Spares Supply (GSS) was established in 2006 in a reorganization of the Service Business Unit (SEB). KONE Global Spares Supply (GSS) is a KONE internal unit that offers and manages over 160.000 different spare parts, maintenance tools, and solutions that are delivered to locations all over the world (KONE 2019). Its mission is achieved by controlling all aspects of the vast materials base. KONE Global Spares Supply main offices are located in Finland and mainland China, and four regional distribution centers (RDC) in Germany, Singapore, China, and Dubai. The main objective of this strategic opening of the RDC is to strengthen KONE's maintenance capability across the region through increased spare parts availability and reduced lead-time. (KONE 2019)

4.1.2 ETL Processes in the Global Spares Supply

The ETL process for moving data from ERP systems to data warehouse is considered as the key integration system for the Global Spares Supply unit. The ERP system, in particular SAP system, is one of the largest global vendors of Enterprise Resource Planning (ERP) software. SAP was founded in 1972 and offers comprehensive solutions for all business processes across finance, supply chain, procurement, manufacturing, service, sales, and human resources. The current successor software to SAP R/3 is known as SAP S/4HANA. With the support of SAP application, the case company could perform their daily operational business activities in a real-time executive environment. In SAP, there are four types of tables including transparent, pool, cluster, and structure table. In this study, the transparent and structure tables are the primary tables in use. The transparent table contains only a single table and is mainly used to store master data. It has a one to one relationship between ABAP Dictionary and the physical

database. Figure 6 shows the technical information of a sample transparent table was captured from SAP.

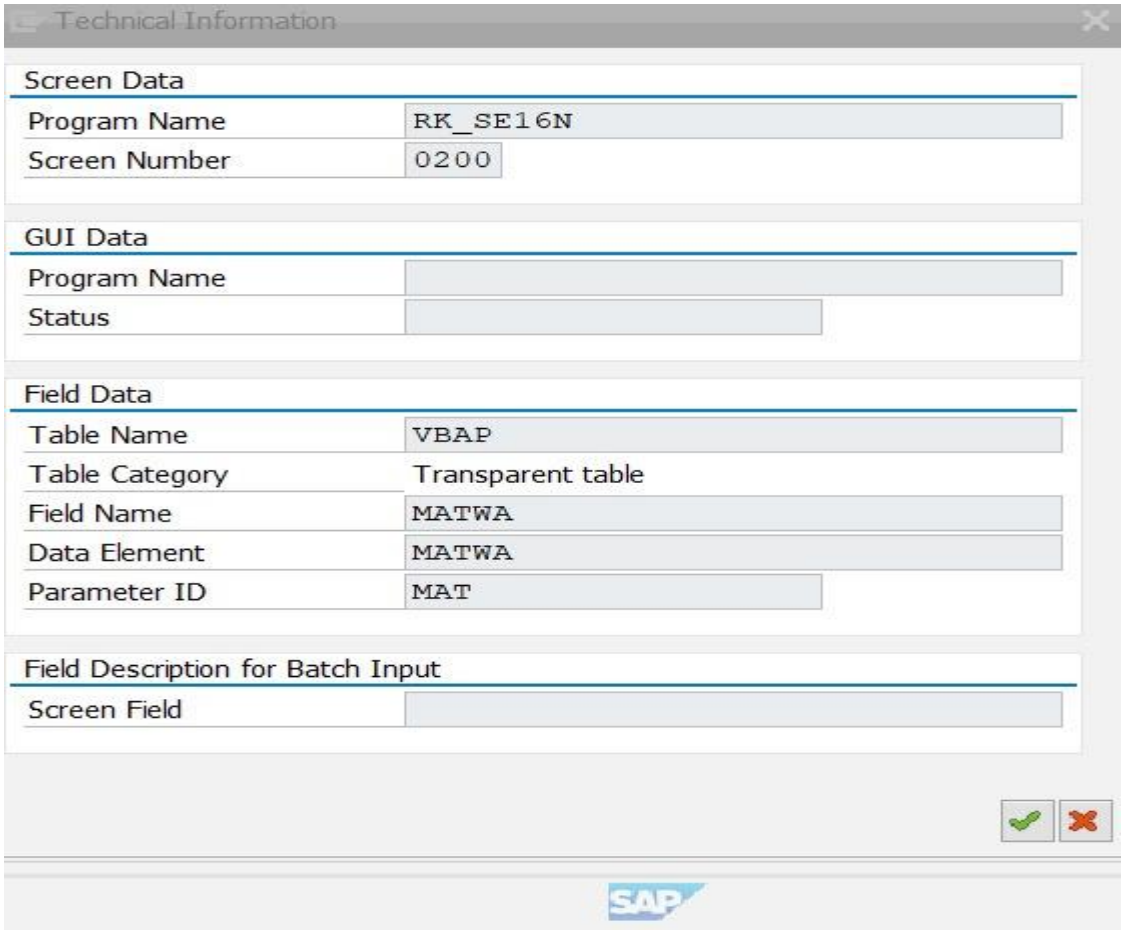

Figure 6. Technical information, the sample transparent table.

Regarding the structure table, they are considered as field strings and no underlying database is generated from them. The technical information of a sample structure table is illustrated in Figure 7.

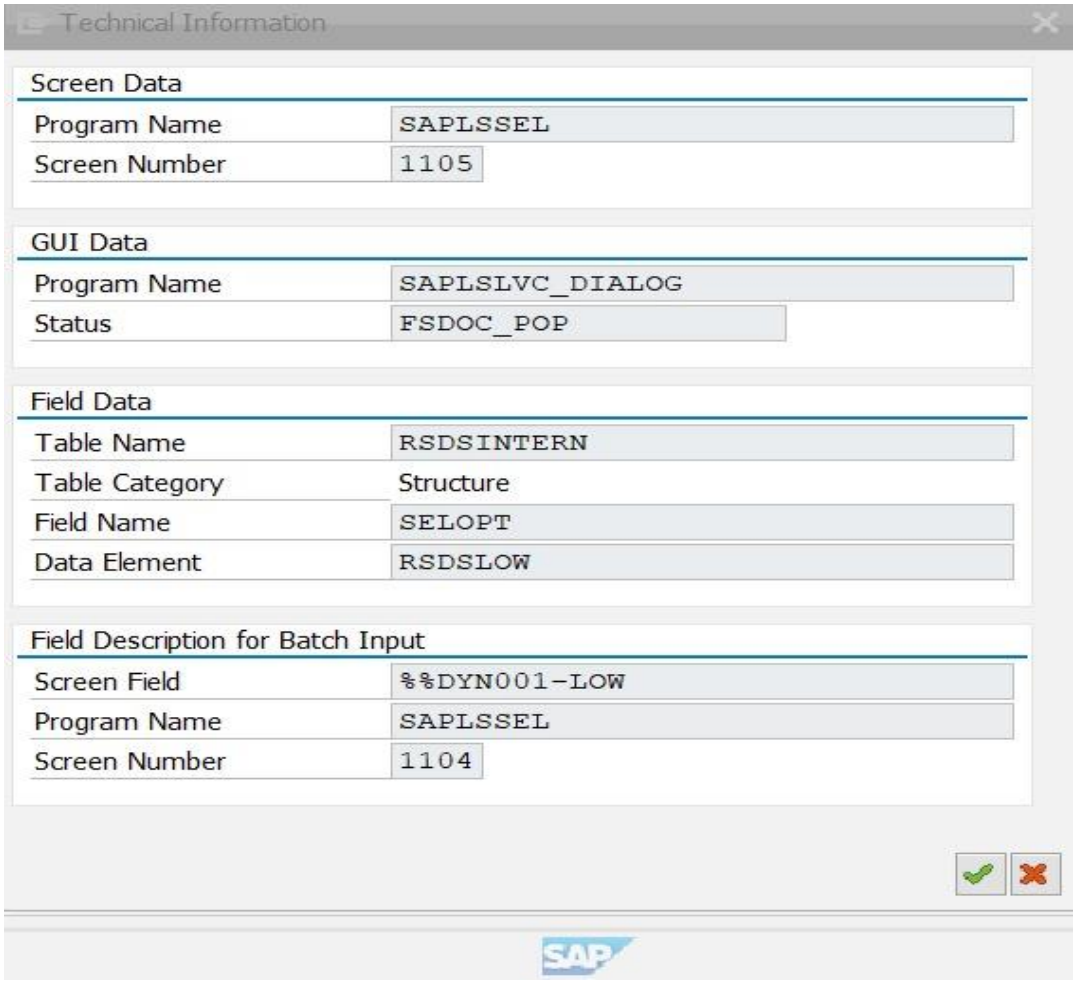

Figure 7. Technical information, the sample structure table.

Data integration, moving data from SAP to MS SQL database is implemented via SAP Connector. SAP connectors allow the integration of different applications and technologies with SAP systems via open standards, the connectors are means for technical interoperability of SAP components (written in ABAP or ABAP Objects) and other components e.g. written in Java, C++, Visual Basic, .NET, C#, etc. (SAP 2020). In this case, the scripting language is used in C#. After setting up the connection between systems, ETL processes are triggered to activate in accordance with the pre-defined schedule. Job scheduling is a tool that provides the ability to schedule the launch of a program or multiple programs or scripts at pre-defined times or after specified time intervals (Wikipedia 2020). Moreover, it is used to wake up a machine remotely or execute a scheduled task. The most commonly used job scheduling tool is Microsoft Task Scheduler. This tool can be understood easily as an event-trigger scheduled task.

A screenshot of the Task Scheduler is included in Figure 8 below. After the task can be created, this event runs automatically at a specific time without human interference.

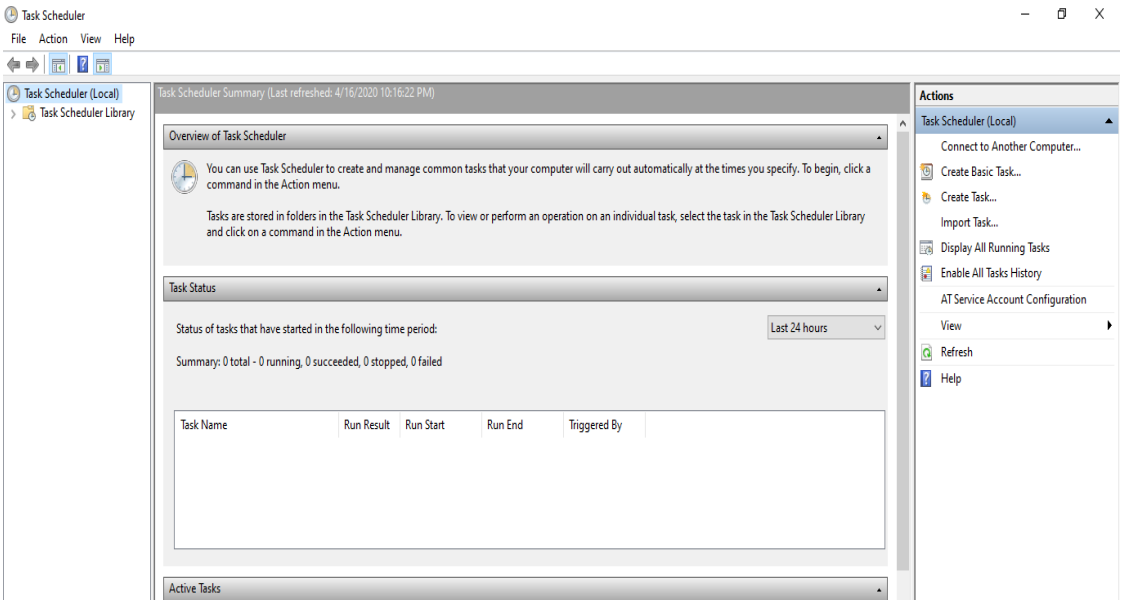

Figure 8. A screenshot of the Task Scheduler application.

After events run at a pre-defined time, the script will produce a log file that will deliver messages status about the result of the query. All log files are collected in a single location on the database server. In computing, a log file is a file that records either events that occur in an operating system or another software run, or messages between different users of communication software (Wikipedia 2020). Besides, an email will be sent to the specified recipients who are in charge of that ETL process.

On the other hand, several SAP structure tables are unable to extract in this way. They are required to access through SAPGUI. SAPGUI is the Graphical User Interface (GUI) client in SAP ERP's 3-tier architecture of database, application server and client, it is software that runs on a Microsoft Windows, Apple Macintosh or Unix desktop, and allows a user to access SAP functionality in SAP applications such as SAP ERP and SAP Business Information Warehouse (now called SAP Business Intelligence) (Wikipedia 2020). This platform is also used for remote access to the SAP server within a company network. The following figure 9 shows SAPGUI interface of the case company.

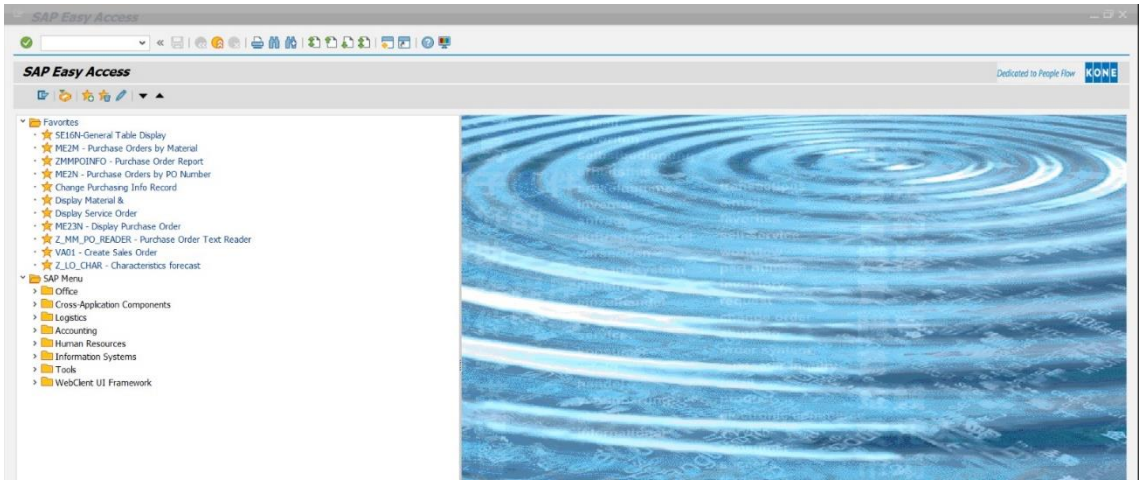

Figure 9. SAP GUI interface.

As for loading data from SAP structure table, the script is written in Visual Basic for Applications (VBA). Visual Basic for Applications (VBA) is an implementation of Microsoft's event-driven programming language Visual Basic 6, which was declared legacy in 2008, and its associated integrated development environment (IDE). (Wikipedia 2020). However, this ETL task is not a complete automation process. Developers must open SAP software and then, activate macro VBA script. From that, data extract activities are executed and data is exported to flat files. Then, data can be imported into data warehouse destination manually.

Integrating system of Microsoft SharePoint that also utilizes C# to load data into data warehouse destination. C# with .NET framework provides powerful libraries for data integration within Microsoft ecosystem.

Besides, data integration tools for website, Outlook email, social media (e.g. tweets, posts, comments) are implemented in R programming language. R is a language and environment for statistical computing and graphics. R packages perform a variety of functions on data processing, manipulation and statistical modeling. Each R script is built dependent upon the website HTML structure and the amount of data needed. Before executing the script, it also builds a connection with Microsoft SQL data warehouse by configuring an ODBC Data Source Administrator. In order to directly build the connection string for ODBC driver to the database, a developer launches ODBC Administrator, creates new data sources and navigates to a target SQL server. The connection string needs to be tested to ensure the configuration run correctly and smoothly. Figure 8 presents ODBC Data Source Administrator.

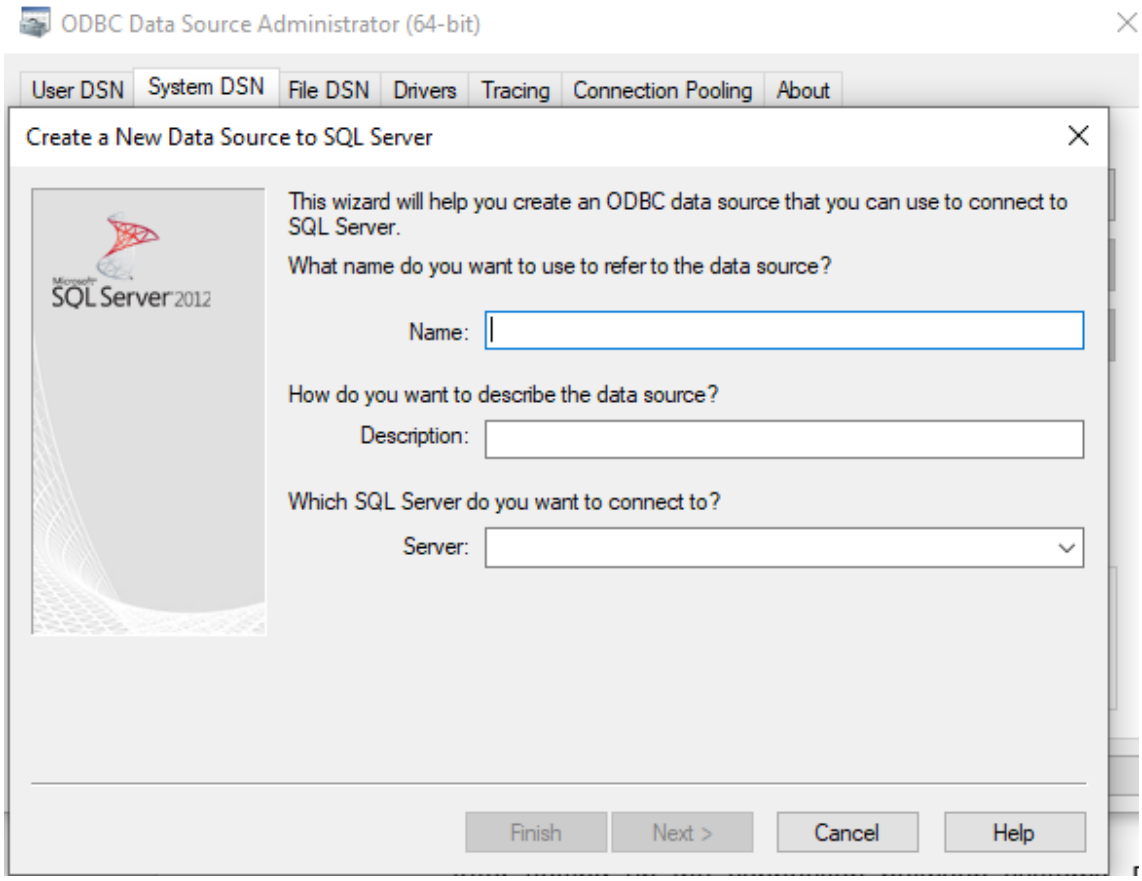

Figure 10. ODBC Data Source Administrator.

Web scraping has become critical to businesses with the rising demand for big data analytics. However, extracting data from multiple websites is not easy at all. There are several challenges, such as complicated and changeable web page structures, IP or user account blocking, dynamic content or complex login requirement, etc. The current R script has still coped with certain difficulties in solving these mentioned issues.

Finally, Figure 11 summarizes the extract, transform, and load (ETL) processes of the Global Spares Supply unit. All these scripts are also attached to the Task Scheduler tool and defined the time to run automatically.

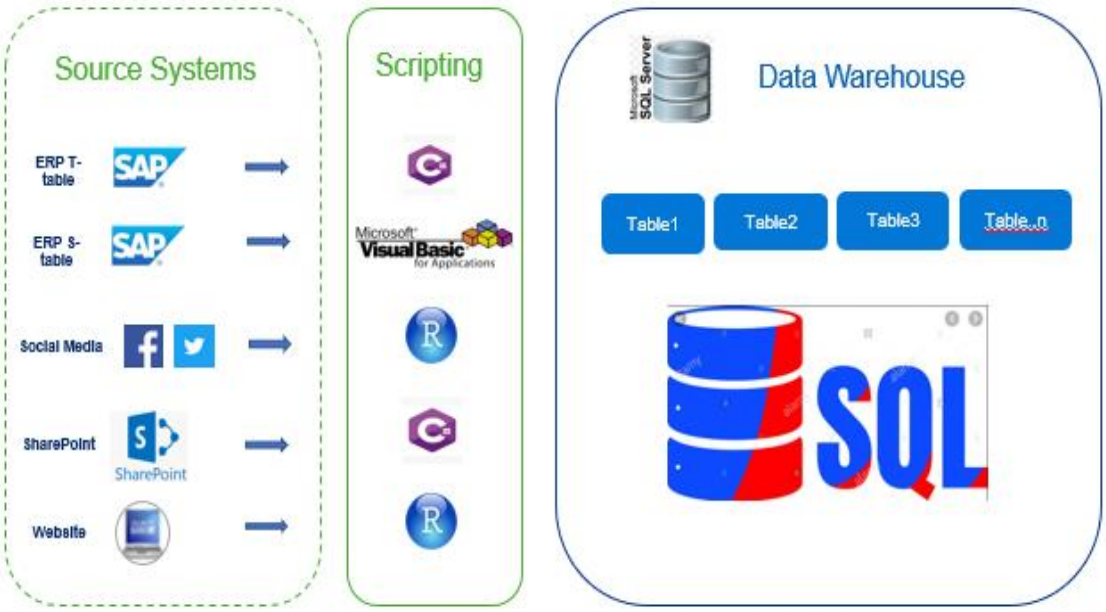

Figure 11. Overview of ETL processes in the Global Spares Supply.

An optimized ETL pipeline processes that allow for efficient migration, stream processing, automated data management, and enrichment of business data. The Global Spares Supply also attempts toward achieving that target. Their expectation will be discussed more in the following part.

4.1.3 Case company needs

The issues of automated ETL processes for several source systems are still a challenge for the case company. There still have manual steps in this process and take a lot of time if this work repeats day-by-day. It is inevitable that human work might cause some certain errors, data security, and especially for time-consuming issues. Managers expect that the current ETL processes can be developed as effective as it could be. There is a desire within the company for more timely information that would facilitate managerial action to be taken. (Rencz 2017). Furthermore, data will be enriched from a variety of source systems that would be critical to understand the market and stay competitive. Thus, building end-to-end automated ETL processes at full capacity for all possible sources would be crucial development at the moment. This study provides related research and appropriate solutions to support the case company.

4.2 Implementation

This chapter describes the implementation of the end-to-end automated ETL processes that are improved with extraction functions comparing to the existing process. These ETL processes are enhanced with full capabilities for all source systems needed at the current time. This implementation consists of three stages, development, testing, and deployment. These ETL developments focused on each source system's requirement. At the current point in time, building an ETL application for both SAP and web scrapping is impossible due to the total difference between source systems' structure.

According to the machine learning and data science survey organized by Kaggle in 2018, Python is the most popular programming language. Indeed, the use of powerful Python libraries not only accelerates the speed of processing huge amount of data, but also supports complex and advanced data structures. For these reasons, this language was chosen as the main programming language in the development. Python is mainly used for extraction, transformation, and load into a target data warehouse automatically as scheduled.

Due to the huge amount of data extracted from multiple sources, there are several infrastructure requirements in order to ensure numerous Python scripts run smoothly. Specifically, processor power should be 4CPUs, RAM is 16GB at least and SSD driver is about 256GB. And the version of Python is 3.6 or 3.7. It would be better to create a virtual Python environment in Anaconda distribution that allows packages to be installed.

There are several Python packages required to install using pip. This is a package manager for Python packages. Besides, the number of libraries can be installed directly on the IDLE (Integrated Development Environment) environment. Table 1 shows the libraries and frameworks applied for every phase.

Phase	**Libraries and frameworks**
Extraction	Win32com, Beautifulsoup, Scrapy
Data discovery	Pandas, Numpy, XML.ETREE, LXML, Matplotlib
Data processing	Re, String, NLTK
Load	Sqlalchemy

Table 1. Frameworks used to handle data.

Thanks to Win32com library, the script can open SAP GUI for Windows system and connect to a specific environment (Z02 or Z04 Production) automatically. Figure 12 shows the source of SAPGUI connection

```
import win32com.client
from win32com.client import *

def Main():

    SapGuiAuto = win32com.client.GetObject("SAPGUI")
    if not type(SapGuiAuto) == win32com.client.CDispatch:
        return
```

Figure 12. Source code of SAPGUI connection.

For user authentication and Single Sign-On in SAPGUI system, user account and password are given in the script. The information about the name of the table and the directory path is also mentioned.

```
session.findById("wnd[0]/usr/txtRSYST-BNAME").text = "user"
session.findById("wnd[0]/usr/pwdRSYST-BCODE").text = "password"
session.findById("wnd[0]/usr/pwdRSYST-BCODE").setFocus()
session.findById("wnd[0]/usr/pwdRSYST-BCODE").caretPosition = 9
session.findById("wnd[0]").sendVKey(0)
session.findById("wnd[0]/tbar[0]/okcd").text = "name_table"
session.findById("wnd[0]/usr/txtP_FNAME").text = "path"
session.findById("wnd[0]/tbar[0]/btn[15]").press()
```

Figure 13. The example source code for SAPGUI authentication session.

The next steps are data discovery and processing. It is dependent on data file formats and structure. For instance, Excel or CSV file format that can be read using the Pandas module. As for complicated XML files, LXML or XML module is very helpful to deal with this data format. Figure 14 shows a simple source code applied XML library.

```
import xml.etree.ElementTree as ET
tree = ET.parse('data.xml')
root = tree.getroot()
```

Figure 14. Example XML code.

After reading data files, data exploration can be implemented with the framework of Exploratory Data Analysis (EDA) which is a visualization approach to summarize data sets and outline the main characteristics. However, there are several different approaches to textual data. Natural Processing Language (NLP) is a typical technique to deal with human languages. This basis of this module provides sentence tokenization, word tokenization, text lemmatization and remove stop words, etc. Besides, other common functions are very common in Python in order to detect and handle missing, null, nan values such as info(), describe(), isnull(), np.nan, replace, re.sub, etc. Data that has been processed and turned into meaningful information. Then, data is transformed into appropriate forms that are ready to load into data warehouse destination.

Before getting into the final step, ODBC string connection must be set up and tested. The screenshot of SQL Server ODBC Data Source Test is demonstrated in Figure 15.

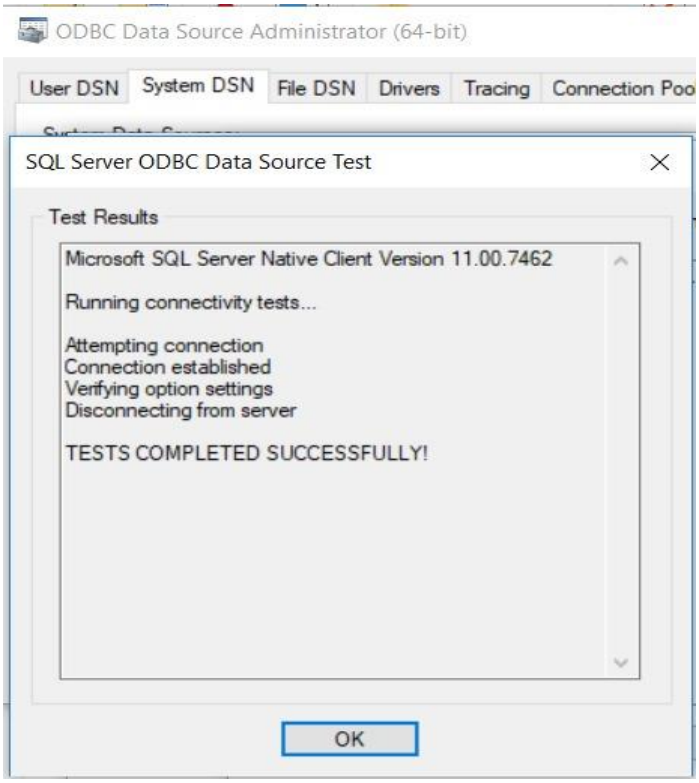

Figure 15. SQL Server ODBC Data Source Test.

Next, SQLalchemy library is utilized in loading data into MS SQL database as Figure 16. User name, password, and database name should be given the query.

```
engine = sqlalchemy.create_engine('mssql+pyodbc://user:password@database')

data.to_sql('name_table', engine,index=False, schema='dbo')
```

Figure 16. The example code of using Sqlalchemy library.

The ETL script for SAP structure table is completed in the development phase. Next, in the testing phase, this will find out and re-check programming work according to the case company requirements and ensure that this script is automated end-to-end. The adjustment can occur in this phase.

After the successful completion of the testing, the script is triggered to run at pre-defined time on Task Scheduler. Specifically, this ETL task will be scheduled and automatized to run every morning at 08:00 AM since April 10, 2020 and forced to stop if it runs longer than 3 days as Figure 17

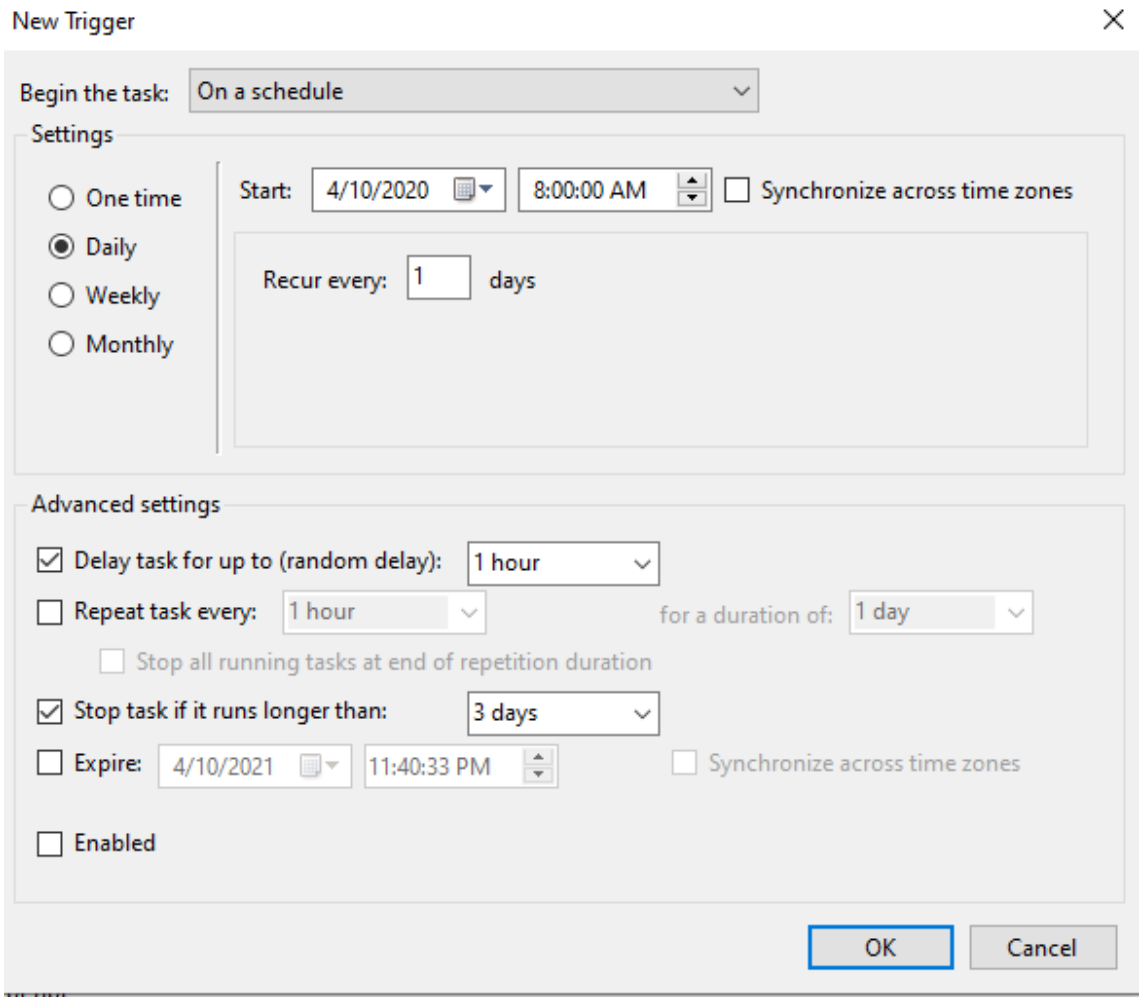

Figure 17. New Trigger on Task Scheduler.

The ETL processes regarding Outlook, websites, social media are also improved thanks to the use of Python programming language. These implementations also include three phases as same as the implementation of the ETL process for SAP structure table. However, the most challenge of these developments is that fetching data from webpages. Not all websites are willing to offer API, therefore, the data retrieval approach is a better way for almost webpages. First, the website needs to be inspected to understand website structure and determine which data can be scraped. Due to the case company policies on confidential information, supplier and competitors' information scraped on webpages is unable to publish. Therefore, Figure 18 shows the structure of an example website that is built for illustration purposes for this thesis.

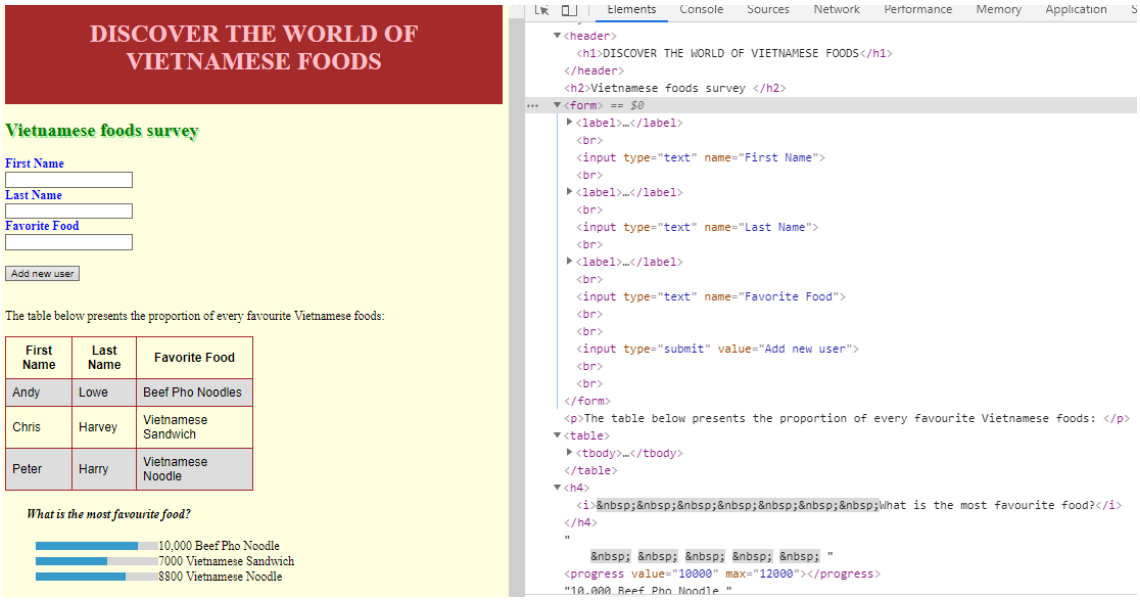

Figure 18. The structure of the example website.

To open the inspector of the web page, please right click on the page and choose "inspect element". Then, it shows the underlying HTML and CSS text of the chosen section of the page. The common libraries (BeautifulSoup, Requests, Scrapy) are called to scrape data. Websites require verification before user authentication. There are several ways to deal with these issues such as POST payload, setting headers, etc. More importantly, the script should add some delay time between events to pretend like a human user.

```
import requests
from bs4 import BeautifulSoup
import pandas as pd
import time

headers = {
        'user-agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_11_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/56.0.2924.8
        'referrer': 'https://google.com',
        'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=0.8',
        'Accept-Encoding': 'gzip, deflate, br',
        'Accept-Language': 'en-US,en;q=0.9',
        'Pragma': 'no-cache'
    }

#User authentication
login = 'login_url'
session = requests.Session()
payload = {'Email': '@mail',
           'Password': 'xxxxxx'}

post_ses = session.post(login, data = payload)

#Scraping
path = 'page'
response = session.get(path, headers=headers)
results_page = BeautifulSoup(response.content,'html.parser')
result = results_page.find('tbody')
rows = resutl.find_all('tr')

for row in rows:
    cols = row.find_all('td')
    cols = [x.text.strip() for x in cols]
    print(cols)

#Delays 5 seconds
time.sleep(5)
```

Figure 19. Example source code of scraping tables on the example web page.

According to Figure 19, there are frameworks prerequisites that must be installed such as Requests, BeautifulSoup, Pandas, and Time. BeautifulSoup from BS4 library is used to parse HTML structure and return the desired information formatted as HTML code. Meanwhile, HTTP POST-Request payload is used for user authentication. Text components and values are scraped from HTML code through the tags and class. Then, data processing and mining techniques are utilized to clean noisy and inconsistent data. Building an ETL pipeline by using BeautifulSoup and Request libraries is successful in the end. The output is illustrated in Figure 20.

	First Name	Last Name	Favorite Food
0	Andy	Lowe	Beef Pho Noodles
1	Chris	Harvey	Vietnamese Sandwich
2	Peter	Harry	Vietnamese Noodle

Figure 20. The output table of the example web page.

The next steps are to connect ODBC string and schedule the scripts to run on Task Scheduler tool on the server.

Thanks to the end-to-end automated ETL pipeline, data can be automatically gone through a button from a web page into data warehouse without human interference. Now, data scraped from the webpage is ready to transform into an appropriate form and load to a target data warehouse. This script is also set to run automatically on a quarterly basis on Task scheduler.

All in all, the improvements contributed to the current ETL processes are summarized in Figure 21.

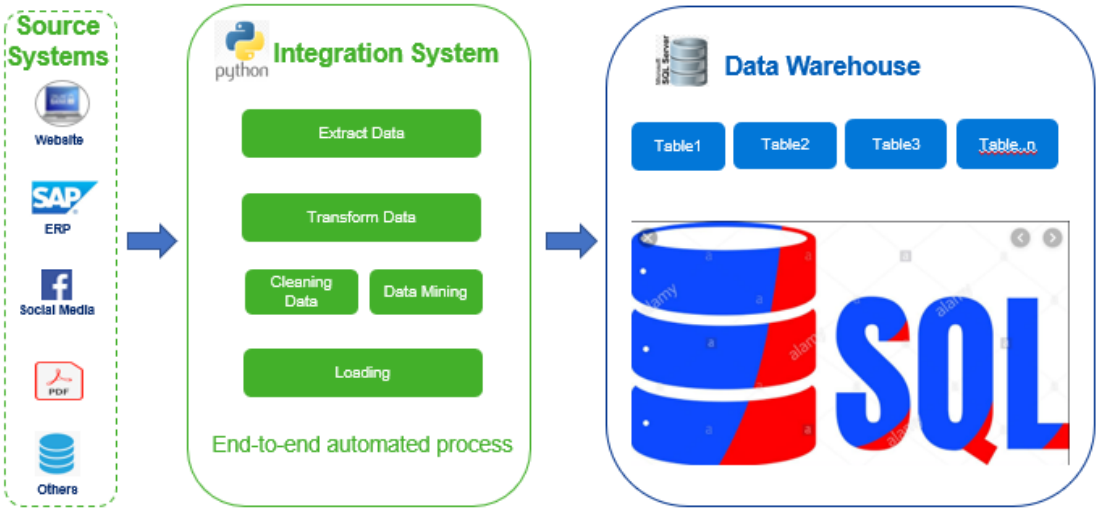

Figure 21. The improvements in the ETL processes.

5 CONCLUSION

In the nutshell, the application of the ETL process is the extracting of large amounts of information on different platforms and loading of a data warehouse. This thesis revealed the importance and benefits of having optimized integration models. This helps companies to manage data efficiently and quickly access data in one data warehouse. With a large amount of historical data stored in one place, companies can obtain a consolidated view that drives better strategic decisions, discovers trends, and make future predictions. Having more data can help with using more powerful analytical methodologies to draw conclusions and create more sales opportunities.

The objective of this thesis was to enhance the current ETL processes of the Global Spares Supply (GSS) unit to increase the performance of the ETL process to full capabilities in extracting data from multiple sources. This study was expected to shift all ETL pipelines to end-to-end automation processes. Therefore, the thesis was action research that focused on both theoretical concepts and real implementation with example code. Multiple challenges were encountered during development phases, the extraction task was most time-consuming.

Finally, the deployment was successful and now ETL processes run automatically on a regular basis on servers. The goal of the thesis was achieved and the ETL scripts were put into production. This development is beneficial to the Global Spares Supply (GSS) unit. The work contributes to the company's needs for timely detailed information and a sufficient amount of data gathered for big data analytics development as well as minimizing misinformation, possible human errors from manual processes.

Despite the importance ETL process, thus far, there are not many researches have been conducted in the area due to its complexity and the differences between companies' data integration processes. Hence, the underlying logic of this thesis can be a basic reference in the early stages of building the ETL pipeline in broader scope in other companies in different industries.

REFERENCES

Aunola, J. (2018) *Data quality in data warehouses.* Master's Thesis in Information and Communications Technology. Lahti: Lahti University of Applied Sciences.

Blichfeldt, B., Andersen, J. (2006). *Creating a wider audience for action research: Learning from case-study research.* Journal of Research Practice, Vol. 2 (1), Article D2. [online] [Accessed 10.1. 2020]
Available at: http://jrp.icaap.org/index.php/jrp/article/view/23/43

Clifton, C. (2010). *Encyclopædia Britannica: Definition of Data Mining* [Online]. [Accessed: 11.1. 2020].
Available at: https://global.britannica.com/technology/data-mining

Cooper, D.R. & Schindler, P.S. 2012. *Business research method*. 12th. New York, McGraw-Hill Higher Education.

Gajare H.P. and Rangdale S. P. 2017. *ETL Data conversion: Extraction, transformation and loading Data conversion*. International Journal Of Engineering And Computer Science ISSN:2319-7242, Vol. 6, Issue 10.10.2017, pp. 22545-22550

Gibbert, M. and Ruigrok, W. (2010) *'The '"What"' and '"How"' of Case Study Rigor: Three Strategies Based on Published Work'*, Organizational Research Methods, Vol.13, pp. 710-737

Han, J. & Kamber, M. & Pei, J. 2006. *Data Mining: Concepts and techniques*. 2nd ed. Amsterdam; London: Elsevier.

Indurkhya, N. & Damerau, F. J. 1. 2010. *Handbook of natural language processing*. Second edition. Boca Raton, Florida: Chapman & Hall/CRC.

Inmon, W.H. 1992. *Building the Data Warehouse*. Hoboken, NJ: John Wiley & Sons.

Kaggle 2018. Kaggle ML & DS Survey. [Online]. [Accessed 1.4.2020]. Available at: https://www.kaggle.com/kaggle/kaggle-survey-2018

Kenneth, R. 2020. *The Hitchhiker's Guide to Python, HTML Scraping*. [Online]. [Accessed 20.03.2020].
Available at: https://docs.python-guide.org/scenarios/scrape/

Kimball, R. and Caserta, J. 2004. *The data warehouse ETL toolkit: practical techniques for extracting, cleaning, conforming, and delivering data,* John Wiley & Sons, Inc.

KONE Oyj. 2019. *Annual Review Kone 2019*. [Online]. [Assessed 10.4.2020].
Available at: https://www.kone.com/en/Images/KONE_2019_Annual_Review_tcm17-88498.pdf

KONE Oyj. 2019 *Enhancing Operational Excellence: KONE Middle East & Africa opens its first Regional Distribution Center (RDC) in Dubai, United Arab Emirates*. [Online]. [Assessed 10.4.2020]. Available at: https://www.kone.ae/en/news-insights/press-releases/kone-middle-middle-east-africa-opens-its-first-regional-distribution-center-in-dubai.aspx

Michel, J. P. 2013. *Web service APIs and libraries*. Chicago, Ill.: ALA Editions.

Microsoft (2020) Introduction to SharePoint Online [Online]. [Accessed 8.1.2020] Available at: https://docs.microsoft.com/en-us/sharepoint/introduction

Mourya, S. K. 2013. *Data mining and data warehousing*. Oxford, England: Alpha Science International Ltd.

Reeve, A. 2013. *Managing Data in Motion: Data Integration Best Practice Techniques and Technologies*, San Francisco: Elsevier, Inc

Rencz, B. (2017). *Enhancing managerial control in production economics through analytics*. Master's Thesis, Hyvinkää: KONE Oyj. GSS.

Rowley, J. 2002. *Using Case Studies in Research*. Management Research News, Vol. 25, pp. 16-27.

SAP (2020) *Data Source.* [Online]. [Accessed 15.3.2020].
Available at: https://www.sap.com/products.html

SAP (2020) *SAP Connector for Microsoft .NET 3.0*. [Online]. [Accessed 10.1.2020] Available at: https://support.sap.com/en/product/connectors/msnet.html

Techopedia (2020) *Data Source*. [Online]. [Accessed 15.3.2020]. Available at: https://www.techopedia.com/definition/30323/data-source

Yin, R. K. 2014. *Case study research: Design and methods*. 5th ed. Los Angeles: SAGE.

Wei, C.C., Liou, T.S. and Lee, K.L., 2008. *An ERP performance measurement framework using a fuzzy integral approach*. Journal of Manufacturing Technology Management, 19, pp. 607-626.

Wikipedia, 2020. *Data cleansing* [Online]. [Accessed: 12.12.2019].
Available at: https://en.wikipedia.org/wiki/Data_cleansing#cite_note-1

Wikipedia. 2020. *Enterprise Resouce Planning* [Online]. [Accessed: 10.3.2020].
Available at: https://en.wikipedia.org/wiki/Enterprise_resource_planning

Wikipedia. 2020. *Log file* [Online]. [Accessed: 3.4.2020].
Available at: https://en.wikipedia.org/wiki/Log_file

Wikipedia. 2020. *Task Scheduler* [Online]. [Accessed: 6.4.2020].
Available at: https://en.wikipedia.org/wiki/Windows_Task_Scheduler

Wikipedia. 2020. *Visual_Basic_for_Applications* [Online]. [Accessed: 15.3.2020].
Available at: https://en.wikipedia.org/wiki/Visual_Basic_for_Applications

APPENDIX A: INTERVIEW QUESTION

1. What is your name?

2. What is your current position at KONE GSS?

3. How long have you been working at KONE GSS?

4. Describe briefly, how does the current ETL processes work? ETL tools?

5. Which data source systems were extracted and stored in data warehouse?

6. What is the business case for every ETL process?

7. What kind of problems or challenges there are, related to the current ETL process?

8. What is your expectation for the improvement of the current ETL tool?

9. What impact does automated ETL bring on your work?