



INTERNATIONAL  
HELLENIC  
UNIVERSITY

# **Data Warehousing in Business Intelligence and ETL Processes**

**Chatzistefanou Dimitra**

SID: 3308200005

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of  
Master of Science (MSc) in Data Science

JANUARY 2023

THESSALONIKI – GREECE



INTERNATIONAL  
HELLENIC  
UNIVERSITY

# **Data Warehousing in Business Intelligence and ETL Processes**

**Chatzistefanou Dimitra**

Supervisor : Prof. Maria Drakaki

**SCHOOL OF SCIENCE & TECHNOLOGY**

A thesis submitted for the degree of  
*Master of Science (MSc) in Data Science*

**JANUARY 2023**

**THESSALONIKI – GREECE**

# Acknowledgment

Major Thanks to my supervisor, Professor Maria Drakaki, for offering supportive guidance and helpful feedback throughout the completion of this Dissertation. Her contribution and encouragement have been invaluable throughout this Thesis .It was my pleasure to have a collaboration with her.

# Abstract

This dissertation was written as a part of the MSc in Data Science at the International Hellenic University. The main purpose is the presentation of the Business Intelligence and Data Warehouse concepts in a theoretical way firstly so as to reflect the important and necessary concepts regarding business intelligence requirements and simultaneously how businesses can respond to the exponentiality of technological changes. Since a Data Warehouse is of vital interest, thorough the core of this Thesis are explained the main definitions, concepts, technical architectures and structured models of Data properties and Data Warehouse in general. It is aimed to offer a comprehensive understanding of the role of data warehousing in business intelligence. Extract, transform, and load (ETL) processes constitute an integral part of Data Warehousing, owing to allowing data to be extracted from a variety of sources, transformed into a standardized format, and loaded into a central repository for analysis. These concepts and knowledge are utilized at the last part of this dissertation by examining a real Dataset from a well-known Czech Republic's bank by employing real world techniques, leading in real world scenarios that constitute useful results that ameliorate business processes and needs.

Chatzistefanou Dimitra

January 2023

# Contents

CONTENTS .....	V
<b>1 INTRODUCTION .....</b>	<b>- 1 -</b>
<b>2 LITERATURE REVIEW .....</b>	<b>- 2 -</b>
2.1 BRIEF HISTORY OF BUSINESS INTELLIGENCE .....	- 3 -
2.2 DEFINITION OF BUSINESS INTELLIGENCE .....	- 3 -
2.3 IMPORTANCE OF BUSINESS INTELLIGENCE.....	- 4 -
2.3.1 METHODOLOGY & LIFE CYCLE OF BUSINESS INTELLIGENCE.....	- 5 -
<b>2.4 DATA WAREHOUSE .....</b>	<b>- 7 -</b>
2.4.1 DEFINITION OF DATA WAREHOUSE .....	- 7 -
2.4.2 DIFFERENCES BETWEEN DATA WAREHOUSE AND DATABASES .....	- 8 -
<b>3 STORAGE ARCHITECTURE OF DATAWAREHOUSE .....</b>	<b>- 9 -</b>
3.1 THE PROCEDURE OF STORING .....	- 9 -
3.1.1 REPLICATION.....	- 10 -
3.1.2 PHYSICAL FRAGMENTATION .....	- 10 -
3.1.3 LOGICAL FRAGMENTATION .....	- 11 -
<b>4 DATA MODELING OF DATA WAREHOUSE .....</b>	<b>- 11 -</b>
4.1 DATA MODELING TECHNIQUES.....	- 11 -
4.1.1 ENTITY-RELATIONSHIP MODEL (ER) .....	- 11 -
4.1.2 DIMENSIONAL-MULTIDIMENSIONAL MODELING.....	- 13 -
4.1.3 OLAP .....	- 14 -
4.3.1 DATA CUBE .....	- 15 -
4.3.2 BASIC OPERATIONS OF OLAP USING DATA CUBE .....	- 16 -
4.4 DATAWAREHOUSE SCHEMA .....	- 16 -
<b>5 GENERAL FRAMEWORK OF TECHNICAL ARCHITECTURE.....</b>	<b>- 17 -</b>
5.1 TOP-DOWN APPROACH.....	- 17 -
5.2 BOTTOM-UP APPROACH .....	- 18 -

5.3 COLLECTIVE APPROACH.....	- 19 -
5.4 TECHNICAL ARCHITECTURE.....	- 19 -
5.6 WORKING PLAN TO BUILD-UP CAPABILITIES .....	- 21 -
<b>6 ETL PROCESSES</b> .....	- 22 -
6.1 OVERVIEW OF ETL IN DATAWAREHOUSE .....	- 22 -
6.2 DATA EXTRACTION .....	- 23 -
6.3 DATA TRANSFORMATION.....	- 23 -
6.4 DATA LOADING .....	- 24 -
6.5 ETL ROLE ON DATA WAREHOUSE & RESPONSIBILITIES .....	- 24 -
<b>7 TECHNICAL IMPLEMENTATIONS OF ETL ON BANKING DATA</b> .....	- 25 -
7.1 THE IMPORTANCE OF APPLYING DATA WAREHOUSE ON BANK INDUSTRY .....	- 26 -
7.2 OVERVIEW OF BUSINESS INTELLIGENCE PROBLEM.....	- 26 -
7.3 OVERVIEW OF USED PLATFORMS .....	- 27 -
7.4 DATA DESCRIPTION.....	- 28 -
7.5 EXAMPLE OF A BI PROBLEM .....	- 28 -
7.6 EXPLANATION AND SOLUTION STEPS.....	- 29 -
7.6.1 ANALYSIS BEFORE SOLUTION WITH IMPLEMENTATION TECHNICAL DW SKILLS.....	- 30 -
7.6.2 SOLVING THE PROBLEM WITH VISUAL STUDIO AND SSMS.- 31 -	
7.7 RESULTS.....	- 38 -
<b>8 CONCLUSIONS</b> .....	- 40 -
<b>9 REFERENCES</b> .....	- 41 -

# 1 INTRODUCTION

Corporations of all sizes and Businesses, for instance banking industry and government agencies, have already realized how pivotal the benefits of implementing state-of-the-art technologies that constantly keep developing to follow the advancements of complex challenges of the modern world are.

Due to the fact that the rate of accelerating technology not only assists businesses to become more competitive and generate revenues and profits but also improves their overall performance, the use of a Data Warehouse in combination with Business Intelligence has become already a necessity.

Data Warehousing is widely recognized as a highly effective method for turning large quantities of data and knowledge within organizations into useful, reliable information and insights. It helps organizations obtain answers to important questions and supports decision-making processes. Furthermore, processes and tools implemented in the context of a Data Warehouse, for instance, ETL (Extract, Transform, and Load), ensure that data is properly formatted for analysis and present time. The overall process of extracting,transforming,loading data ,known as ETL,plays a fundamental role in maintaining the accuracy and quality of the data stored in a data warehouse.

Consequently, in the modern world, data warehousing and BI are becoming increasingly important as organizations aim to take advantage of the large amounts of data generated through their operations and external sources. Data warehousing and business intelligence assist businesses to identify opportunities for growth and cost savings and gain a comprehensive view of their data.

## 2 LITERATURE REVIEW

Nowadays more than ever, no one can deny the extensive importance of data-driven solutions and state-of-the-art technologies applied in today's digital world, and that Data Warehouse is a critical component of Business Intelligence since it offers the central repository for data. Hence, it is crucial to cultivate a holistic idea of Business Intelligence (B.I.) and how Data Warehouse serves B.I. requirements for gaining a more complete and accurate view of the organization's data. In this section, it is pointed a literature review concerning Business Intelligence concepts and Data Warehouses.

Through the writing of this thesis, definitions concepts, models, and architecture implemented on actual Data Warehouse concepts are defined by scientific papers, written by distinguished academicians and practitioners, to frame data warehouse effectively in a comprehensive way.

The literature concerning Business Intelligence and Data Warehouse implemented ETL processes, is broad and it covers many topics. As the preliminary concepts for Business Intelligence and Data Warehouse are fully illustrated and explained combined with technical implementation concepts and models to cultivate a general understanding of the data-driven concepts implemented. Traditionally, a significant portion of research and discussion in the field of data warehouses has focused on the area of conceptual modeling and design. This includes both academic literature and the research community.

All in all, a review of the literature shows that data warehousing is a key component of BI since it results in the creation of a centralized repository of data that can be queried and analyzed to support business intelligence activities.

## 2.1 BRIEF HISTORY OF BUSINESS INTELLIGENCE

**Duplicity Practised by Furness, the King's Banker.**

The name of Sir Henry Furness figures largely among the bygone bankers who gave renown to the financiers of that period. Throughout Holland, Flanders, France, and Germany, he maintained a complete and perfect train of business intelligence. The news of the many battles fought was thus received first by him, and the fall of Namur added to his profits, owing to his early receipt of the news. On another occasion he was presented by King William with a diamond ring of immense value, as a reward for some important information, and as a testimony of that monarch's esteem. He was the king's friend and banker.

But the temptation to deceive was too great, even for this eminent and honored banker. He fabricated news; he insinuated false intelligence; he was the originator of some of those plans which at a later period were managed with so much effect by Rothschild. Thus, if Sir Henry wished to buy, his brokers were ordered to look gloomy and mysterious, hint at important news, and after a time sell. His movements were closely watched—the contagion would spread; the speculators become alarmed; prices be lowered four or five per cent.,—for in those days the loss of a battle might be the loss of a crown,—and Sir Henry would reap the benefit by employing different brokers to pur-

Figure 1

The first emergence of the term “Business Intelligence” (abbreviation: BI) was introduced before the advent of computers. It initially appeared in the book ‘The Cyclopedia of Commercial and Business Anecdotes’, in 1865, written by Richard Miller Devens [FIGURE 1 & 2]. He introduced this term when one human character in his book, Mr. Sir Henry Furness, who was a banker, started collecting abundant information and details about customers and the market and based on This concrete evidence he was led to the capability to respond and make business decisions leading to increased profit. As this constitutes the main concept of BI, the banker

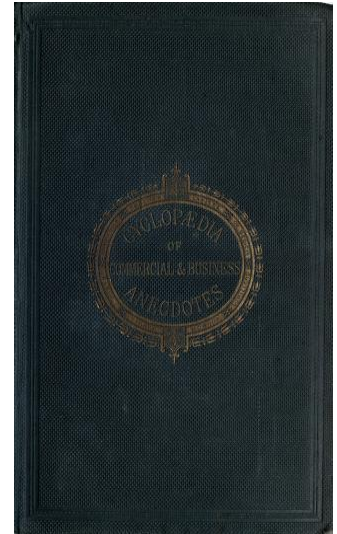


Figure 2

Furness assisted his bank corporation to become more competitive towards other banks and competitors obviously by conducting better, more efficient, and sustainable actions that lead to refined business impacts. Of course, currently due to the rapid advents of technology, business intelligence methods are more improved than in the 19th century, using Business Intelligence as an umbrella term for many technological advances regarding business decisions and decision support systems, that keep improving nowadays on the market of Technology with the collaboration of terms of ‘Big Data’ and ‘Data Mining [1] [2] [3].

## 2.2 DEFINITION OF BUSINESS INTELLIGENCE

In paper of H.P Luhn [4] there is a clear definition of the term Business Intelligence, by separately analyzing each word of this term. The notion of ‘Business’ is depicted as the gathering of all activities and processes perform for every sector of society, for instance for governments, organizations, technology, law, science, etc. The word, ‘Intelligence’ is characterized the overall capability and activity of comprehending the associations of facts and events that are currently happening and afterward taking into consideration these linkages to conduct actions that aim to the wanted objectives.

In other words, Business Intelligence is considered a combination of processes and practices with an analytical, and technology scope that transfer collected data and every type of information into useful and powerful business insights that undoubtedly serve enterprises' objectives and in general the positions of every enterprise and organization. By taking advantage of analyzing data - both recent and historic- intends to refine decision-making so as to frame more actionable the strategic behaviors and provide meaningful insights.

With the term Business Intelligence, the system used combines collection gathering from multiple sources, storage, and knowledge management. One of the major goals is to simplify the data in a way to be acknowledged by managers and business analysts and to be turned in a more readable and understandable form. This process of conversion procedure needs a diversity of actions that should be followed. The data processing initially after the extraction of the data requires the data preparation that included transformation, cleaning, and filtering of data and afterward the following the step is the loading of data to a storage area – a repository. This step need of a repository leads to the notion of Datawarehouse. [5]

## **2.3 IMPORTANCE OF BUSINESS INTELLIGENCE**

It cannot be denied that Business Intelligence is a critical tool for big enterprises and organizations for many reasons .The most important reasons are the following :

- make better decisions [6]
- increase efficiency and ameliorate streamline processes [6]
- gain a competitive advantage and differentiate themselves from competitors [7]
- improve customer satisfaction [6]
- stay compliant [6]

## 2.3.1 METHODOLOGY & LIFE CYCLE OF BUSINESS INTELLIGENCE

The methodology that it is followed for Business Intelligence is developed by implemented the stages depicted in [Figure 3] [8].

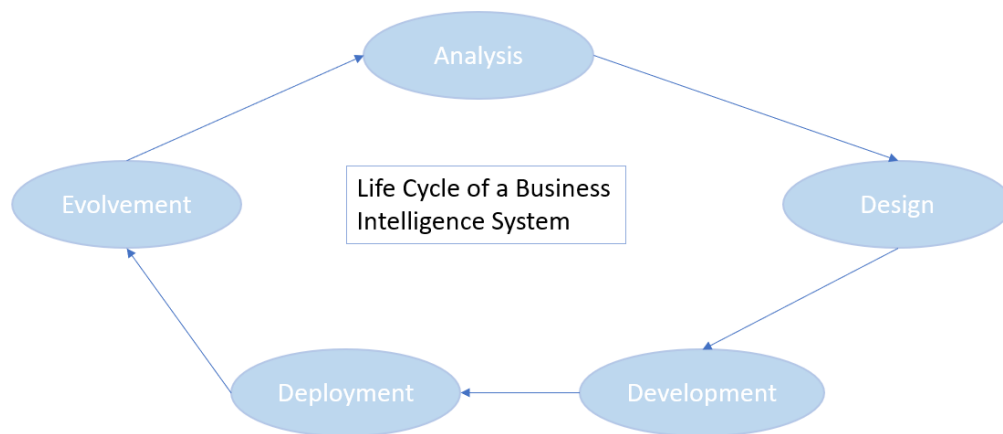


Figure 3

### 2.3.1.1 ANALYSIS

The first phase of undertaking to solve a Business Intelligence problem should be characterizing of the proving and analyzing the parameters existing included the cost and the advantages of solving this problem

### **2.3.1.2 DESIGN**

In this step of Business Intelligence life cycle , the technologies that should be utilized have to be chosen so as to be suitable to the problem in terms of sophistication of the problem .The planning and designing has to follow a model which simulate the process of solving the business problem and finally to be led to the final solution. Hence ,dependent to this blueprint the adequate resources and steps and business methods and decisions have to determined ,by visualizing the delivery prerequisites and the prospects.

### **2.3.1.3 DEVELOPEMENT**

In this step of development, technical implementations and the database model that have to be applied is required to suite properly to the model. ETL (stands for Extract, Transform and Load) processes are often used to the Development stages and also Data mining techniques are used for supporting this development stage. Data mining is an umbrella term that mention that generate predictions , meaningful insights and aims to firm performance and uses algorithms and techniques such as Classification , Clustering , Nearest Neighbor method ( NN ) , Decision Tree , Linear Regression and Non-Linear regression .

### **2.3.1.4 DEPLOYMENT**

At this step , the previous development step has to successfully finished and tested . With the intention to ensure high quality , testing and adjustments are applied from end users so as to meet business needs. As a consequence , this interactions between users and data are a step for future optimization of data.

### **2.3.1.5 EVOLUTION**

After all the previous stages and successful executions of technical implementations that lie correctly to enterprise needs , the application is implemented to the enterprise and gets continually tested and reported , so as to give valuable impact for future modifications .

All in all , these five steps of Business Life cycle are repeated every time with tuning every time parameters that optimize enterprise profits each time.

## 2.4 DATA WAREHOUSE

As it was mentioned above, Business Intelligence is one of the most crucial aspects for every company and organization so as to grow well and gain more profit in comparison to their competitors. The notion of a Data Warehouse is included in one of the most important activities of Business Intelligence. Sometimes the notion of a Data Warehouse and its connection with Business Intelligence is ambiguous. In other words, a data warehouse constitutes a component of Business Intelligence. That is the clear relationship between these two notions.

### 2.4.1 DEFINITION OF DATA WAREHOUSE

The term and definition of data warehouse was first introduced by an American computer scientist , Bill Inmon.

Four are the main characteristics of the collection of data, as DW is defined, encouraging strategic decision-making procedures and giving worthwhile business insights . These characteristics are depicted in the following image [Figure 4] :

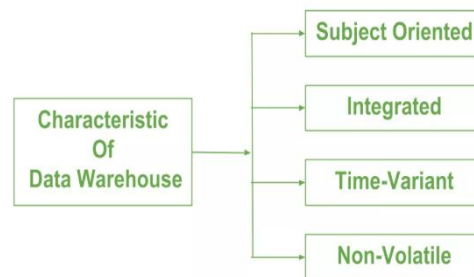


Figure 4

Data warehousing is the actions taken to well organize data in a systematic way so as to have powerful and insightful knowledge through mining processes. These four characteristics describe Data Warehouse The definition is explained below:

1. Subject-oriented: Data is organized by being categorized around fundamental subject areas like sales, claims, loans, transfers, letters of guarantee, Accounting, shipments, and enrollments. For instance, client transfers knowledge contains all the historical records of transfers of each client. over specific time intervals. The data are organized in a way that has connection and relation to real-world events and the objects are closely connected This principle specifies that complete subject

information should be stored in a single place. The subject should not consist of any extra data that are not relevant. A data warehouse delivers information on certain subjects such as procurement, products, marketing, sales, etc. Therefore, data could be archived based on a particular team, as opposed to the current operation conducted by a firm.

2. Integrated: The fact that a data warehouse is created by integrating and combining data from different sources, such as a mainframe, relation databases, etc, results in the characterization of a Data Warehouse as "integrated" [9].
3. Time-variant: The data is stored through different time intervals such as weeks, months, quarters, and years. Data warehouse aims at the storage and organization of data with information and analytical processing using an extended range of historical time.
4. Non-volatile : Seeing that data within a data warehouse is historical data, it is categorized as permanent and could not be erased, edited, or deleted. Changes to data which it is stored in A data warehouse environment is prohibited, unless it is occurred in a controlled and organized way – in contrast to the OLTP environment. The update of data occurs only when the next scheduled load takes place. Consequently, changes are not reflected in data stored in Data Warehouse (data remain stable and consistent before the next scheduled load.)

As a consequence, to conclude the definition of Data Warehouse is that subject-oriented points to the analysis of any subject, time-variant and integrated refers to the retainment of historical data from different sources and is non-volatile to the content of a Data Warehouse that cannot be altered [10] [11] [12].

## **2.4.2 DIFFERENCES BETWEEN DATA WAREHOUSE AND DATABASES**

Both databases and data warehouses constitute systems specially designed for storing data, but they have different purposes. Individuals often get confused about the notions of data warehouses and databases because they both store data. It is essential to distinguish the key differences between them. Source () In DB, short-term and isolated operations are organized and repeated, while the storage capacity needed is hundreds of megabytes or gigabytes. On the other hand, Datawarehouse addresses various needs and is built to extract data from multiple sources, model and store the data in a single location, and use OLAP to provide a multidimensional view of the data. A data warehouse aims to consolidate data and information under one data structure model, such as in a decision support system. In addition, concerning the data used in the Database versus the Data Warehouse, many are the number of key differences, but the most important ones that must be highlighted is [13] [14] [12]:

1. Optimization: Data warehouses are designed to be optimized for analysis, while databases are optimized for transactions
2. The orientation of data: Theme-oriented is the data used in the Data warehouse, while the data of the Database is application-oriented
3. Update of data: Because DW is non-volatile, data cannot be updated, compared with DB that there is this capability.
4. Speed of performance: In cases of data complexity, DW performs operations slowly, while the performance speed on DB is immense.
5. Data warehouse data is summarized and at times improved, while in DB the form of data is intricate.

## **3 STORAGE ARCHITECTURE OF DATAWAREHOUSE**

A key technical component of a data warehouse implementation is Data Storage. Storage architecture is a fundamental element.

The storage component of the data warehouse serves as the foundation of the system and is where all the data is kept and available for use. The creation of a consolidated and Integrated view of the organization constitutes the grand organization's objective, where various concepts are semantically linked.

### **3.1 THE PROCEDURE OF STORING**

Storage is an essential component of a data warehouse because it is the location where all the data will be kept and then transformed to be utilized by the organization. In addition, security reasons and also ease regarding the overall view and understanding of the concepts and parts of DW and the way that different concepts are connected [15]. The procedure for storing a DW can be seen in the following image [Figure 5]:

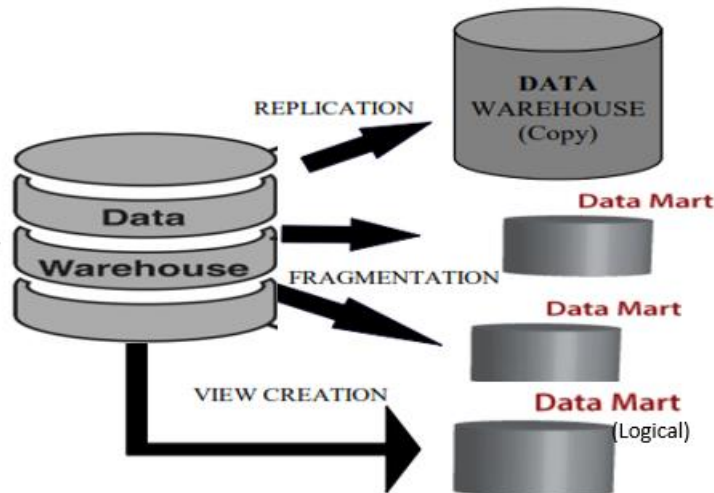


Figure 5

The three stages are the following: Replication , Physical fragmentation, Logical fragmentation.

### 3.1.1 REPLICATION

Keeping a data backup is critical for the constituency of a Data Warehouse environment for many reasons but mainly for security issues [Inmon, 1996b], . Also keeping an updated replicant in cases of corruption of the system or failure or loss of data [16].

### 3.1.2 PHYSICAL FRAGMENTATION

Physical fragmentation. this procedure which is non-demanding but optional is the use of data marts.

Data Mart constitutes a smaller version of a Data Warehouse which only deals with a single subject, and it is considered a subset of corporate-wide data that's useful to a selected group of users. Data Marts are specified for one area.

Hence, data marts are designed for data that come from a restricted number of sources. Also, it can be easily concluded that comparing the time taken to build Data Marts and a Datawarehouse, the construction of a Data Mart needs less time than a Data Warehouse [15].

### **3.1.3 LOGICAL FRAGMENTATION**

The impact of the logical fragmentation is considered to be the identical to the physical , except for the creation of a new data mart or a storage area .In other words , Logical fragmentation offers an overall overview of enterprise's data without the need of any other actions , like data transformation or data movement [15].

## **4 DATA MODELING OF DATA WAREHOUSE**

Data modeling constitutes a way to create a representation of the real world in order to better understand and analyze it. This visualization helps to see things that may not be immediately apparent and aids in the creation of plans and strategies. Essentially, data modeling allows us to model and understand complex systems in a simpler and more intuitive way.

The data model serves as a blueprint for implementing a data warehouse, providing a clear and structured plan for how the data will be organized and used. It helps to ensure that the data warehouse is designed in a way that is efficient, flexible, and conducive to fast querying and analysis [17].

### **4.1 DATA MODELING TECHNIQUES**

In the field of data warehousing, it is common to use two different techniques for data modeling : 1)Entity-Relationship modeling (ER) 2)Dimensional Modeling

#### **4.1.1 ENTITY-RELATIONSHIP MODEL (ER)**

ER modeling is a method for constructing a data model of a specific domain by identifying the entities within it and the associations among those entities. Additionally, this approach also involves specifying attributes, which are properties of either the

entities or the relationships. This technique is helpful for understanding and simplifying complex data relationships in business and system environments. The entity-Relationship (ER) model is depicted using an ER diagram, which uses three main graphical symbols to visualize the data: entities, relationships, and attributes.

#### **4.1.1.1 ENTITY**

Entity: A person, place, thing, or event of interest to a business or organization that represents a class of objects that can be identified and classified based on their characteristics and properties.

#### **4.1.1.2 RELATIONSHIP**

Relationship: The structural interaction and association between entities, depicted by lines in a model and represented grammatically by verbs such as 'owns', 'belong', and 'has'. The association between two entities is characterized by cardinality, which denotes the maximum quantity of instances of one entity that are linked to a single instance of the other entity (such as one-to-one, one-to-many, or many-to-many).

Firstly, One-to-one mapping involves linking each element in the target data warehouse to a single element in the source data warehouse (for instance a table, record, or attribute) corresponds to a single element in the source. Secondly, in many-to-one mappings, an element in the target data warehouse is derived from multiple elements in the source. Finally, in many-to-many, multiple elements in the source are combined to create multiple elements in the target data warehouse [10].

#### **4.1.1.3 ATTRIBUTES**

Attributes: The properties or characteristics of an entity that are described in the model are known as attributes. Attribute names should be unique within an entity and self-explanatory and may be nullable or optional if their value is not specified for a particular instance. When the maximum number of instances of an attribute is more than 1, the data modeler may choose to split the attribute into a separate entity to increase data organization and efficiency [18].

### **4.1.2 DIMENSIONAL-MULTIDIMENSIONAL MODELING**

On the other hand, Dimensional modeling is based on concepts: facts, dimensions and measures. It is useful for representing the needs of business users in the context of database tables.

Dimensional- Multidimensional modeling is a widely used design approach for constructing data warehouses [19].

Dimensional modeling is mainly used to support Online Analytical Processing (abbreviation :OLAP) and decision-making processes. To understand the business requirements of OLAP and model a data warehouse effectively, it is useful to review some basic concepts of OLAP. There are four main operations used in OLAP to analyze data: drill down, roll up, slice& dice. OLAP operations allow us to analyze data at different levels of detail and navigate through dimensions. The examination and explanation of OLAP will be discussed next.

Multidimensional modeling is a valuable method for illustrating and representing data models as a group of measurements that possess various business attributes. This technique is useful for organizing and presenting data in a way that supports analysis, while numerical data, including quantities, amounts, weights, balances, and frequencies, can be effectively analyzed. Dimensional modeling involves several key concepts, including facts, dimensions, and measures (variables) [18].

#### **4.1.2.1 FACTS**

Fact: A group of related data items that include both measures and context data, representing a business item, transaction, or event that can be used for analyzing business processes. In a data warehouse, these facts are stored in core tables that contain numeric data.

#### **4.1.2.2 DIMENSIONS**

A dimension is a group of members or units of the same type that provide context for the facts in a data model. In a multidimensional model, each entry in the fact table corresponds to a single member from every dimension. These dimensions, which represent the parameters for online analytical processing (OLAP), may include time, location, customers, and salesperson, among others [18].

#### **4.1.2.3 MEASURES**

Measure: A numeric element of a fact that reflects the operation or activity of a business in relation to the dimensions. Measures, which may include sales, volume, quantity, cost, and transaction amount, are determined by the combination of dimension members and are located on facts. In a relational context, there are two main models used in dimensional modeling [18].

#### **4.1.3 OLAP**

The use of The Multidimensional Model is a highly effective way of conceptualizing data for use in On-Line Analytical Processing (OLAP) applications [20].

Many are the reasons that the multidimensional model offers beneficial and valuable impacts on Data Warehouse projects, but mainly for the ease of analysis: analyze and compare data from different perspectives while simultaneously it aligns with the way analysts think, and it can significantly improve query performance. As a consequence, it constitutes valuable aid for decision-makers [20] and an intuitive way for business people to analyze data in a more natural and easy way.

To conceptualize data for use in On-Line Analytical Processing (OLAP) applications is the effective approach of The Multidimensional Model [20].

OLAP (Online analytical processing) is frequently utilized in data warehousing and business intelligence due to its capability to allow users to manipulate and analyze data from various sources. It constitutes a key set of systems and techniques that a data warehouse engineer uses to answer important business intelligence questions through the analysis of data stored in the data warehouse. OLAP enables users to efficiently perform advanced analytical queries on large datasets and utilizes data modeling and aggregation techniques. As a result, OLAP helps decision-makers gain valuable insights and make informed business decisions [21].

Additionally, online analytical processing (OLAP) systems are designed to arrange and display data in a variety of formats to accommodate the varied requirements of various users. These systems allow users to analyze and manipulate data from multiple sources in a quick and easy manner, using advanced data modeling and aggregation techniques [22].

#### 4.1.4 DATA CUBE

A data warehouse's effectiveness in analyzing and summarizing data can be improved by using a multidimensional model. This model conceives data as residing within an n-dimensional space, often called a data cube or hypercube [23].

A data cube, depicted in [Figure 6], is a multidimensional model in the shape of a cube, where each cell represents a set of values for various dimensions. As a multidimensional data structure is utilized to store and examine substantial amounts of information.

A data warehouse is constructed by fundamental cuboids, which are surrounded by other cuboids that show how the base cuboids are combined in various dimensions. Overall, a data warehouse with d dimensions will have  $3/4 d$  sub-cuboids [24]

Dimensions as mentioned previously on 4.1.2.2 are the entities which an organization – in our occasion the banking wants to keep records.is a set of smaller attendant tables , one for each dimension. simply defined, a dimension is a categorically consistency view of data .

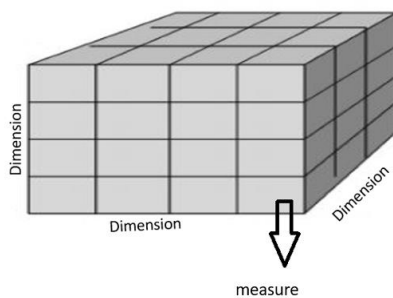


Figure 6

To make better decisions faster , managers require useful information that is readably and easily understood so as to be can make changes on them and keep track of data. LAP constitute an important tool for relation data and supports the refinement of the decision support procedure [25].

Once all the cuboids have been calculated, it is easy to use OLAP operations like roll-up, drill-down, and slice-and-dice within the graph cube framework .

#### **4.1.5 BASIC OPERATIONS OF OLAP USING DATA CUBE**

Some common OLAP operations performed on a data "cube" include aggregating data (roll-up), exploring data in more detail (drill-down), selecting specific data (slice and dice), and rotating data (pivoting) [26].

The roll-up operation involves synthesizing data by moving from a lower level to a higher level in a dimension hierarchy or by decreasing the number of dimensions. In contrast, the drill-down operation involves decreasing the level of aggregation or increasing the level of detail, which can be achieved by adding new dimensions. Both the roll-up and drill-down operations can be applied to the elements of a dimension hierarchy [26].

The slice and dice operation involves selecting a subset of the data's dimensions and projecting the data onto those dimensions while fixing the values of the other dimensions. This reduces the number of dimensions in the data [27]

An instance of this could be to select customers from a certain age group in the context of a business case of finding a target marketing group . or customers by a certain as income level or geographic location.

The pivot operation involves changing the orientation of the data in a cube through rearrangement [26].

All in all, OLAP operations using the data-cube give the opportunity to end users to interactively analyze and explore processes using a multidimensional view of event data.

## **4.2 DATAWAREHOUSE SCHEMA**

A schema is a logical structure that describes the entire database, including the names and details of all record types and their associated data items and aggregate values while specifying the manner that data is organized and stored within a database.

Star schema: A star schema consists of a central table containing the core data , referred as the "fact table" , is connected to a series of additional tables that provide context and additional information, known as "dimension tables".The fact table holds the main data that is being analyzed, while the dimension tables hold descriptive data about the entities being analyzed. The star schema is ease of use, and has a intuitive structure that is unchallenging to understand and query.

Snowflake schema A snowflake schema is a type of database design that is like-minded to a star schema, but the dimension tables are organized in a more complex way since they are normalized and divided into multiple tables, creating a "snowflake" shape.This

structure allows for more complex relationships between data elements but can be more difficult to understand and query.

Like snowflake schema, is also a table schema that supports multidimension schema. A fact constellation schema (also known as a multi-star schema) is a structure that is commonly used in data warehousing. It consists of a principal fact table that is connected to multiple dimension tables [19].

In conclusion, Data modeling involves creating a blueprint for how data will be organized and used within a specific application. Over the last decades, data modeling has been a significant focus of research and development in various fields, such as databases, information systems, software engineering, and knowledge representation [28] [29].

## **5 GENERAL FRAMEWORK OF TECHNICAL ARCHITECTURE**

Enterprise architecture (EA) is a comprehensive representation of organizations that encompasses various aspects of business operations such as terminology, organizational structure, processes, and data, as well as automation elements like information systems and databases, and technological infrastructure including computers, operating systems, and networks. Thus, Enterprise architecture serves as a cohesive force that aligns objectives, vision, strategy, and governance principles and facilitates collaboration between these elements [30].

The architecture of a data warehouse pertains to the overall layout and organization of a data warehouse. There are several different approaches to designing the architecture of a data warehouse, including the top-down, bottom-up, and a combination of both approaches, called collective approach.

### **5.1 TOP-DOWN APPROACH**

The top-down approach concerns starting with a general plan and preparation. It is effective within organizations where expertise is well-established and the business issues to be addressed are clearly understood.

In the top-down approach to implementing a data warehouse, a significant amount of planning and design is completed at the beginning of the project. This involves the participation and involvement of individuals from many different businesses parts. Prior to implementation, decisions regarding data sources, security, data structure,

data quality, data standards, and the overall data model are often made. Furthermore, that approach, that starts from a higher level and works down toward the details, can result in a more uniform understanding of data and the implementation of company policies throughout the entire organization. However, the initial planning and design costs can be substantial and may delay the implementation and benefits of the data warehouse. This approach is generally suitable for corporations with a centralized information systems organization (IS organization) responsible for all computing resources but may not be suitable for organizations with decentralized IS resources or workgroups, departments, or lines of business with their own IS resources. Moreover, the top-down approach is often utilized to create a global data warehouse, with data marts being built and populated from the global data warehouse rather than directly from operational or external data sources [17] [31] .

## **5.2 BOTTOM-UP APPROACH**

The bottom-up approach involves starting with small-scale models and gradually building up to larger ones. This approach is suitable during the early stages of business development, when equipment and resources may be limited. It enables organizations to move forward with minimal upfront costs, allowing them to assess the potential benefits of the technology before committing significant resources to its implementation.

In contrast to a top-down approach, a bottom-up implementation of a data warehouse concerns the planning and design of data marts without waiting for a global infrastructure to be put in place. This approach has become more and more popular in recent times since it allows for immediate results from data marts, which can be used to justify the expansion of a global implementation. Furthermore, data marts, which are smaller and focused subsets of a larger data warehouse, can be constructed independently or simultaneously with a global data warehouse. These marts can be populated with data either from the global data warehouse or directly from operational or external sources. One significant advantage of the bottom-up approach is that it gives the opportunity for faster impacts due to the less complex design of data marts and lower initial implementation costs. However, there are also some potential drawbacks to consider, such as the potential for data redundancy and inconsistency between data marts, increased load on operational systems, and difficulties integrating data marts into a global environment. Overall, it is important to carefully contemplate all these factors before choosing the bottom-up approach for implementing a data warehouse.

## 5.3 COLLECTIVE APPROACH

The utilization of both a top-down and bottom-up method in the execution of a data warehouse each possess their own advantages and disadvantages. In many scenarios, utilizing a blend of both strategies may be the optimal solution. The combined approach allows organizations to use the detailed planning of the top-down approach while still being able to take advantage of the quick execution of the bottom-up approach. This provides the ability to utilize a well-thought-out plan while still being able to take advantage of opportunities as they arise.

However, this combination can be challenging, but with a skilled project manager, it is possible to achieve a balance between the two approaches. One key to success is to determine the level of planning and design that is required for the global data warehouse to support integration with data marts built using the bottom-up approach. Firstly, identify the lines of business that will be participating and develop a high-level view of their business processes and data areas of interest. When creating data marts, it is important to establish a strategy for addressing data elements that are required by multiple data marts. This could involve creating a common data store or duplicating data across multiple data marts. Moreover, careful monitoring of the implementation process and management of issues can maximize the benefits derived from both approaches.

## 5.4 TECHNICAL ARCHITECTURE

The General The framework that Technical Architecture addresses the following topics for overall architecture and implementation practices for EDW. In addition, the Integration The architecture chapter addresses the following processes for batch data processing where transformation and standardization are required at the record level or across sets of records:

- Data acquisition that involves obtaining data from internal or external sources for data integration and storage in operational and data warehouse repositories.
- Data Harmonization ensures consistency within the acquired subject area data.
- Data Population involves loading data into and providing data within and across data repositories.

### **Data Acquisition**

Data Acquisition is a critical process that enables the data value chain to function properly. It constitutes the procedure of collecting, filtering, and cleansing data before the procedure of storing it in a Data Warehouse [32].

## Data Harmonization

As the world becomes progressively dependent on tech standards and data information technology, the need of employing more data harmonization approaches for ensuring compliance with a variety of data conditions is rapidly rising. Since the sources of the acquisition of data keep increasing, the form of data is literally characterized as heterogeneous and decentralized. The diverse nature of data that have different formats and follows dissimilar structures such as semi-structured, and structured. It is quite challenging to find ways to achieve data harmonization so as to be led to high-quality data, and as a consequence to provide decision-makers with the right information and data [33].

Data Harmonization constitutes the whole process of resolving inconsistencies between related datasets in order to bring their structure and content into an agreement. Common data structures and the content will be defined per Subject Area and Feed. These common structures enable the integration of data throughout the system.

Data harmonization involves Validation, Normalization, and Transformation processes. Data Harmonization relies upon standardized Reference and Lookup data to support Validation and Transformation. Error/Exception Handling processes, for instance [ Figure 7 ] will be used in the event that a record fails to validate or transform. Audit, Balance, and Control procedures are used to ensure establish and maintain data integrity.

Once data has been “Harmonized” into a common structure and content, it can be made available for use in downstream processes, including Data Enrichment and Data Population.

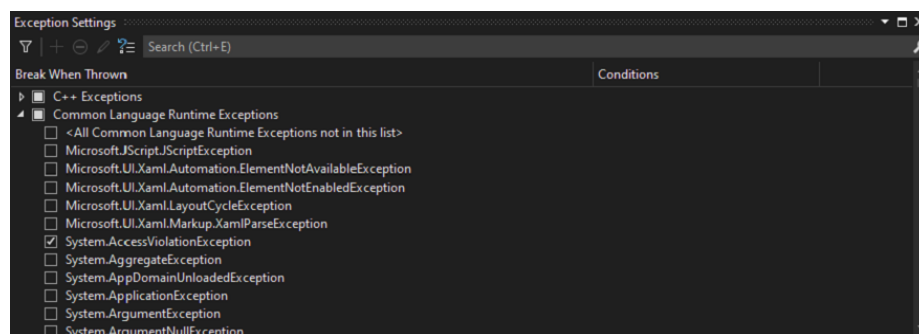


Figure 7

## **Data population**

The Data population concerns procedures, methods, and techniques so as to populate data to tables of Data Warehouse. Demands, that are defined by businesses and corporations, in combination with Service Level Agreements (SLA) specify the overall and suitable table load strategies considering the data volumes. More precisely, the notion of SLA can be described as follows:

A service level agreement (SLA) constitutes a contract commitment between a service provider (for instance a bank) and the customer-client that frames the specific type of service that the customer desires to acquire and the actions-measures should be followed in particular cases and business circumstances. In other words, SLAs typically cover quality of service, service availability, and provider responsibilities [34].

It is worth mentioning that data population involves: Loading strategy, slowly changing dimensions, common table loading, and aggregation.

For instance, concerning slowly changing dimensions, a slowly changing dimension can be regulated in three different manners: Type 1, Type 2 and Type 3. It is worth elaborating on each type.

Firstly, in type 1 the new values totally replace the original value, without keeping any historical record or trace of previous values.

Secondly, in type 2 approach does not overwrite the existing value but appends the new value with the goal to keep track of previous records and become historical. Every time that a new record occurs, at dimension table is updated by the addition of one new row having a new identifier code for every alteration. For this particular reason, Type 2 approach, in some specific circumstances, the identification principle is absent due to the existence of more than one identifier code.

Moreover, type 3 : Attempting to avoid Type 2 lack of integrity, in Type 3 there is the addition of a new record attribute as a unique row having the same code as an individual identifier. In this case of employing the Type 3 approach, there is a limitation of the number of changes that cannot exceed a specific number of alterations [35].

## **5.5 WORKING PLAN TO BUILD-UP CAPABILITIES**

Analysis : The design phase of the IT DW projects .Make clear Business demands and goals , functional specifications and attempt to depict business requirements on technical framing [36].

Implementation: Technical Specification . Develop and test the tools . Document the process [36].

User Acceptance Test (U.A.T.) : During this phase, the integration and testing of business and technical processes is carried out. Additionally, the data is evaluated for its accuracy, completeness and quality. Business units must also accept the results of the technical procedures.

Generally , a User Acceptance Test is defined as a guarantee of customer satisfaction that verifies that contractual obligations are met. Especially , concerning data warehouse context , A UAT involves testing a data warehouse to confirm that it satisfies the needs of the users, such as its ability to store and analyze necessary data, provide accurate and current information, and support the intended business processes and decisions [37] [38].

Deployment : Deploying a data warehouse involves setting up the system in a production environment and testing it using real data. If any corrective action is necessary, adjustments are made to improve both processes and implementation time [39].

Go live : The bank will have the capability to operate on a live basis on an as-needed (at-hoc) basis, monthly, or daily within a specific time frame [40].

## **6 ETL PROCESSES**

### **6.1 OVERVIEW OF ETL IN DATAWAREHOUSE**

ETL (stands for extract, transform, and load) constitute an essential and valuable tool that is being applied to the data warehousing process. In data warehousing, ETL refers to the overall procedure of employing specialized software to gather data from various sources, manipulate it in a staging area, and then upload it into a data warehouse. ETL tools are utilized so as to facilitate this process. Additionally ,these tools are specialized software that are used to gather data from various sources, clean and customize it, reformat it, integrate it, and load it into a data warehouse Creating an ETL (Extract, Transform and Load) procedure constitutes a significant task when developing a data warehouse. It is a complex and resource-intensive task that often consumes a significant amount of time, effort, and resources [41].

The ETL process ,illustrated on [Figure 8] ,typically follows the framework shown in [Figure 8].Information is gathered from multiple sources and moved to a designated

location for data storage (DSA). In the DSA, the data is processed and cleaned before it is added to the Data Warehouse. ETL tools are used to manage and automate this process [41].

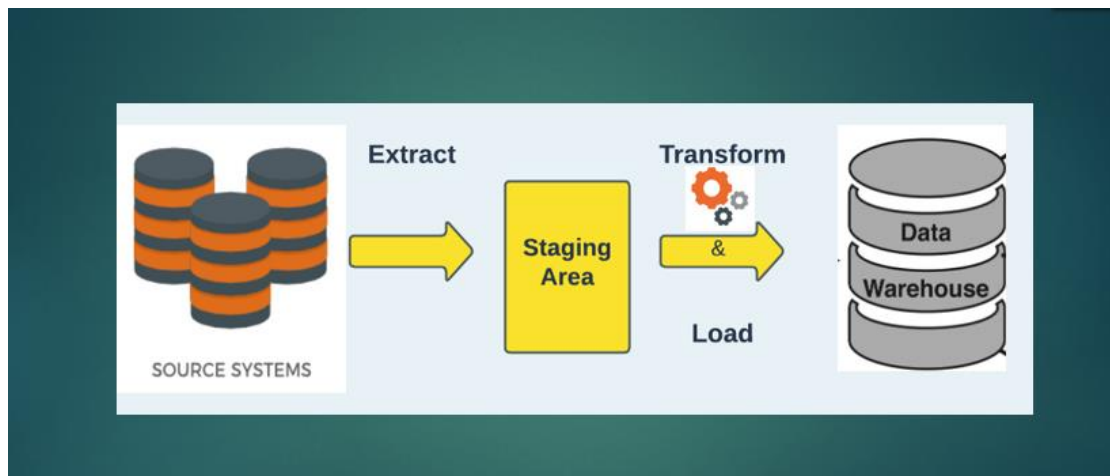


Figure 8

### 6.1.2 DATA EXTRACTION

**Extract :** Data extraction is the initial and also a significant step in the ETL process. In the extract phase, data is extracted from multiple and disparate sources, such as databases, files, or other systems. The extracted data is frequently disordered and can come in a range of formats and arrangements.

There are two types of data extraction: full extraction, which involves extracting all of the data from the source system, and incremental extraction, which involves only extracting changes made to the source system. One example of a tool that uses incremental extraction is Change Data Capture (CDC) [42].

### 6.1.2 DATA TRANSFORMATION

**Transformation:** During the data transformation phase of the ETL process, data is converted to be stored in a format or structure that is suitable for querying and analysis. This procedure involves tasks such as data cleansing, reformatting, matching, aggregation, generating surrogate key values, splitting columns into multiple columns, sorting data, converting coded values to their corresponding meanings or labels, and structuring free-form data with the intention to prepare the data for storage in the data warehouse [42] [43].

#### **6.1.4 DATA LOADING**

Load: In the load phase, the transformed data is transferred to the final destination, often a data warehouse (DW), where it may replace or be added to existing information. This typically involves creating tables or other data structures in the data warehouse and inserting the transformed data into these structures. The DW is often updated on a regular schedule, such as hourly, daily, weekly, or monthly, using the extracted data [42].

During this process, the data produced in the transformation process is applied to the appropriate data warehouse or data mart, as previously mentioned. There are four principal techniques for loading data: load(insert), append(add), constructive merge, and destructive merge. The load method replaces the current data in the target data warehouse tables with the data from the transformation process and can also create the target tables if they do not already exist. The append method adds the new data from the transformation file or table to the end of the existing data in the target table. The constructive merge method combines the new records that are created with the already existing data in the Data Warehouse table and simultaneously updates an end-time value for any superseded records. The destructive merge technique completely replaces the current records with the new data.. The smooth operation of a data warehouse is fundamentally dependent on ETL. The destructive merge technique completely overwrites the existing records with the new data [17].

Overall, the ETL procedures are essential for the correct operation of a data warehouse. If the source data from various sources is not processed correctly, the data warehouse will not be able to provide strategic information, and the querying process that forms the foundation of the data warehouse will be definitely incorrect and misleading. It is critical for ensuring that the data warehouse can support efficient and effective data analysis [44].

### **6.2 ETL ROLE ON DATA WAREHOUSE & RESPONSIBILITIES**

The role of the ETL (extract, transform, load) architect is vital in designing and implementing the ETL process for a business intelligence system. It is the responsibility of the architect to ensure that the process meets the needs and requirements of the organization and to work with technical staff and infrastructure architects to develop an efficient method for extracting and transforming data. Additionally, it is important for the ETL architect to ensure that the technical team understands the target database design and the transformations required to convert

source data into the target data structures. In addition to these responsibilities, the ETL architect serves as a central point for understanding technical standards such as data design standards, file size limits, and requirements for passing data through firewalls. The ETL design process should be repeatable, well-documented, and subject to proper change control. The ETL architect may also serve as a consultant to the programming effort and may supervise the development of programming specifications. Moreover, it must guarantee the selection the appropriate software tools for different types of data processing and the recognition of the differences between the design and implementation methods for a business intelligence system and an online transaction processing system. The ETL process is crucial for maintaining and sustaining a business intelligence architecture and strategy [45].

## **7 TECHNICAL IMPLEMENTATIONS OF ETL ON BANKING DATA**

With the intention to understand in a clearer way the procedure and steps ETL process, one small representation of ETL will be an illustration of a real Data Warehouse. This illustration employs a real dataset of loans by the Czech Data released for 1999 PKDD's Discovery Challenge, which provides access freely to real development data such as account information, loan, and credit card records.

This dataset consists of financial records derived from a Czech Bank and includes approximately 5300 clients with over 1 million transactions. Also, the data of this dataset include around 700 loan data and information about 900 credit cards [46].

As mentioned above, the dataset used in this Thesis is a real dataset that concerns financial details, and simultaneously it delivers information about clients,

transactions, loan details, etc. of a Czech Republic bank. Thus, The Berka Dataset is a real-world dataset that only extremely sensitive clients' information (for example names, surnames, and ID numbers) is anonymized for customer privacy-preserving reasons.

It is well recognized that the Berka dataset is used as the standard for measuring the efficiency of financial data mining and fraud detection techniques. It is commonly employed in research and academic environments. Since financial data are highly confidential files, it is remarkably

In general, to have access to real banking data is considered very challenging. Overall, the Berka financial dataset is a valuable resource for researchers interested in analyzing financial data.

## **7.1 THE IMPORTANCE OF APPLYING DATA WAREHOUSE ON BANK INDUSTRY**

The banking sector is known for being a data-intensive and technology-driven industry due to a large amount of information it handles and the rapid advancements in technology. As the number of banking operations increases, there is a growing need for implementing business intelligence, which is the process of collecting, managing, and analyzing large amounts of data from sources such as customers, products, and transactions. However, the constant improvement of technology and the increasing volume of data can make it difficult to efficiently handle all this information using only business intelligence techniques. One solution to this problem is the use of data warehouses, which transform data into valuable insights and can be exclusively constructed to serve the banking's system's needs [47].

## **7.2 OVERVIEW OF BUSINESS INTELLIGENCE PROBLEM**

In the context of this Thesis, it is created a BI exercise project, which is developed and materialized by construction of a piece of DW techniques. The stimulation of a real DW is created by Microsoft Technology platforms, more precisely with Visual Studio - Integration Services and Analysis Services- and SQL Management Studio – an integrated environment by Microsoft for managing support SQL infrastructure and in general SQL operations.

## 7.3 OVERVIEW OF USED PLATFORMS

One popular tool that developers use to create a variety of applications, including



Figure 11

those for data warehousing, is Visual Studio [Figure 11]. This integrated development environment (IDE) offers a wide range of features and tools for designing, developing, and deploying data warehousing solutions. It includes tools for data modeling, ETL processes, and data visualization [48].

Microsoft developed a programming toolkit called Visual Studio in 1997. It includes a code editor, debugger, and designer, and allows developers to create console and GUI applications for various Microsoft platforms using languages like C#, C/C++, and Visual Basic. The code editor has a background compilation feature that highlights errors in red, and the debugger enables developers to find and fix bugs in their code at both the machine and source levels. Visual Studio also has extensibility options. It offers a set of tools and features that handle data warehousing solutions, including tools for data modeling, ETL (extract, transform, load) operations, and data visualization. [48]

In addition, SQL Server Management Studio (SSMS) [ Figure 12] is also widespread tool that is used for managing and developing data warehousing solutions.

SSMS offers improved performance and expanded capabilities . Its features provide powerful insights for businesses and also enhance the performance of online



Figure 12

transactional processing (OLTP) and data warehouse solutions [49]. SSIS is also considered as a tool used to perform ETL processes ( the whole procedure of extracting, transforming, and loading ). The system is engineered in a way that it can quickly process data in memory row by row and then transfer it to the database structure of a data warehouse. [50].

Some of the capabilities that SSMS provides are the creation of databases, schemas, and tables, writing and executing SQL queries, and offering feedback in cases where a query does not follow the required rules so as to be correct. Therefore , The Query Editor is a popular development tool in SSMS that allows users to write and execute Transact-SQL (T-SQL) queries, view results and create execution plans [51].

Furthermore, it allows the designing and deployment of ETL processes and also the visualization of data with the employment of a wide range of graphical tools and reports. Moreover, SSMS is principally utilized in data warehousing projects by developers since its inclusive features and tools are specialized for being used for Data Warehouse projects.

## 7.4 DATA DESCRIPTION

This data set consists of 8 different files, and each file is considered as different information for each client.

The raw files that exist give information about accounts, clients, disposition, permanent order, transactions, loans, credit cards, demographic data.

Every account has some static characteristics (stable values that they are not affected and changing) and some dynamic characteristics ( values that are updated so as to fit to the current image of each account) . To be more precise , static are the permanent values that do not change for every client (for instance, date of birth, loan details about the day that the loan payments started etc.) . In contrast, dynamic characteristics are the fields that their value change and reassigned due to transactions and other banking procedures ( for instance : every client that perform a transaction , their balances is altered [52].

The analysis of every raw file that Berka dataset provides is the following :

- ACCOUNTS file: Each record contains details about the unchanging attributes of a particular account.". (Size: 4500 objects in the file)
- CLIENTS file: Every record contains information about a client's characteristics. Each record describes the unique client that connects with an account. Every account may have more than clients connected with this
- DISPOSITION (DISP) file: record links a client with an account, for instance, records describing the client's rights to operate the account (Size: 5369 objects in the file)
- PERMANENT ORDERS, Debits only (ORDER) file: Every record includes information about the characteristics of a payment order. (Size: 6471 objects in the file)
- TRANSACTIONS (TRANS) file: Each the record describes one transaction on an account (Size: 1056320 objects in the file)
- LOANS file: Each record provides details about a loan granted to a specific account. (Size: 682 objects in the file)
- CREDIT CARDS file: Each the record describes a credit card issued to an account

## 7.5 EXAMPLE OF A BI PROBLEM

It is widely accepted that one of the biggest Business Intelligence goals is the understanding of clients-customers in a way that decisions should be taken about the improvement of profits-earnings may be data-driven. In General, the fact is that Enterprises are searching for ways to constantly improve their services and get some

extra profit by making their clients consume more bank products (cards, prepaid cards, loans, etc.)

The business intelligence problem exercise that in this chapter will be performed is a customer segmentation problem.

In general customer segmentation is considered an intensive valuable strategy for big corporations and businesses that aim to market their products and develop new ones more effectively. This practice results in dividing the market into distinct groups based on characteristics such as needs, characteristics, or behavior, which may require unique products or respond differently to marketing efforts. This can help companies to tailor their marketing efforts more precisely and increase the effectiveness of their campaigns. In conclusion, customer segmentation leads to identifying dissimilar segments of customers according to their attributes, such as their age, gender, occupation, and financial behavior [53].

In particular, the business problem is to identify the clients that pay their loans on time (Making timely payments) or they have completed the cycle of payments of the series of installments-payments. These customers are considered “good” or “performing” because they meet the terms on their loan agreement.

In terms of solving our Business Intelligence problem is to create a table on the data warehouse that includes only the “performing” customers and discriminates them from the “non-performing” ones with the intention to send them messages (emails or messages or letters at their home address etc.) with informing content that concerns the product that fit them.

A significant problem that also emerges is to find every time the right clients and to send them messages (emails or messages or letters at their home addresses) with informing content that concerns the product that fits them but, in a way, not to be spammed. Clients should be informed only one time. So, it is crucial to keep and compare with historic data so as to inform always new clients.

All things considered, the ultimate goal is to create a DW table that contains only the ‘performing’ clients-customers who have fully consistent in paying their loan since the business goal is to seek clients who are fully consistent to their already-paid loans, so as to provide them with the right bank-products that fit them and find them.

## **7.6 EXPLANATION AND SOLUTION STEPS**

An important point in solving the problem is that every time this table is loaded, there should be no duplicate entries, i.e., there should be no customers recorded in this particular Data Warehouse table more than once. In cases where enterprises send many marketing messages, it is considered a violation of laws and regulations regarding unsolicited communication. As a consequence, every time that the list of the “performing” clients is needed, it should not contain clients that have been already informed about the marketing products.

The improvement that managers need to achieve is to make ‘good’ customers are pleased with no complaints about being disturbed by advertising messages with new products that they do not need. To illustrate this, if clients cannot meet the expense of paying their loan debt, it is certain that they are not interested in ‘purchasing’ new bank products that will make them become more in debt.

### 7.6.1 ANALYSIS BEFORE SOLUTION WITH IMPLEMENTATION TECHNICAL DW SKILLS

Our business intelligence problem requires only the file of loans. For the purpose of understanding better our dataset that is given, it is following an analysis of loan dataset (Figure [13] ) .

The file of loans includes the following information: loan\_id is record identifier, account\_id is Account Identifier, date is the date when

	A	B	C	D	E	F	G
1	loan_id	account_id	date	amount	duration	payments	status
2	5314	1787	930705	96396	12	8033.00	B
3	5316	1801	930711	165960	36	4610.00	A
4	6863	9188	930728	127080	60	2118.00	A
5	5325	1843	930803	105804	36	2939.00	A
6	7240	11013	930906	274740	60	4579.00	A
7	6687	8261	930913	87840	24	3660.00	A
8	7284	11265	930915	52788	12	4399.00	A
9	6111	5428	930924	174744	24	7281.00	B
10	7235	10973	931013	154416	48	3217.00	A
11	5997	4894	931104	117024	24	4876.00	A
12	7121	10364	931110	21924	36	609.00	A
13	6077	5270	931122	79608	24	3317.00	A
14	6228	6034	931201	464520	60	7742.00	B
15	6356	6701	931208	95400	36	2650.00	A
16	5523	2705	931208	93888	36	2608.00	A
17	6456	7123	931209	47016	12	3918.00	A
18	7104	10320	931213	259740	60	4329.00	A
19	6820	9034	931216	38148	12	3179.00	A

Figure 13

loan was first granted (a form of date is year-month-day: YYMMDD, the amount is Amount of the loan, duration is the duration of the loan , status is the loan’s paying off condition.

In loan table there is a field, named Status with 4 different values: A , B , C , D

- A: Contract finished, and loan is fully paid with no problems.
- B: Contract finished but the loan is not paid so far.
- C: Running contract, with no problems.
- D: Running contract but client in debt and do not pay.

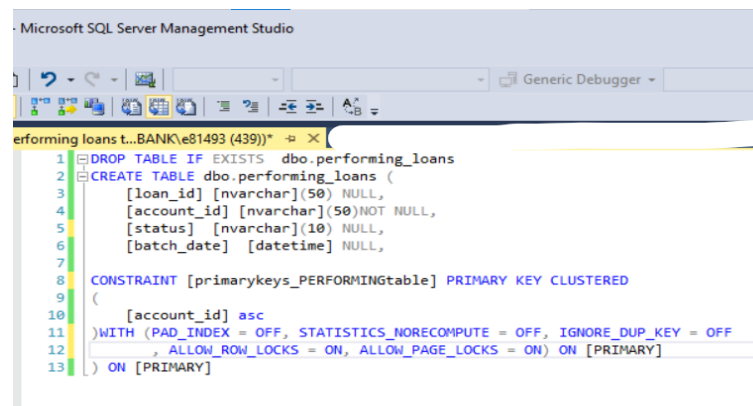
The business goal is to find the clients who are fully consistent to their already paid loans, so as to provide them with the right bank-products that fit them and find them. On account of this, we are interested only for clients in cases A and C. In case that clients (in case C) fully pay their loan, their status will be A. Since loan transaction data is updated and not static (like other static characteristics of an account like account number) .

As time goes by, some clients in ‘C’ category that they fully pay their loan, they will become category ‘A’. As a result, we are going to choose only the clients where their loan status is either “A” or “C”. There, the business intelligence problem using Data

Warehouse practices will inform only one time a client even if the client turns from status “C” to status “A”.

## 7.6.2 SOLVING THE PROBLEM WITH VISUAL STUDIO AND SSMS

In SSMS, it is created a table named “dbo.performing\_loans” ( dbo: is the name of the schema used to create this table ) [Figure 14].

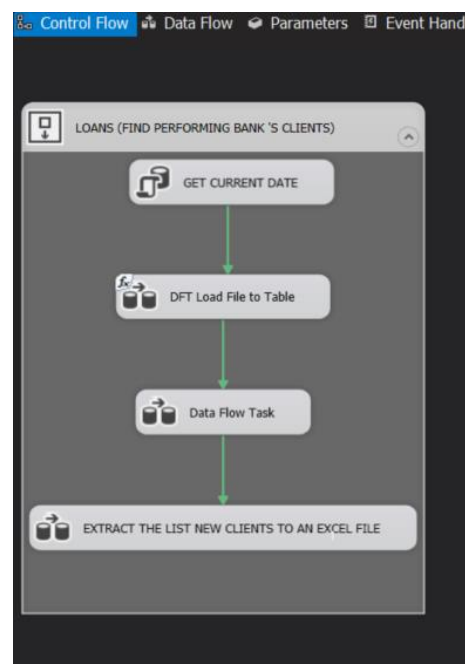


At this specific table , the data with clients who have shown a good conduction when it comes with their loan payment.

Figure 14

We have the loan\_id and account\_id fields as they exist on table dbo.loans .

The table dbo.loans contains all the records that raw file loans of Berka dataset



include. Afterwards, in Visual Studio platform now, the initial step here that has to be taken is to take the current date . First component of the package gets the value of current datetime , which is accomplished with the command “GETDATE()” and then convert the datetime to date with the format “YYYY-MM-DD” year, month, and day) so as to have the exact date that the procedure that it is followed. That is achieved with the T-SQL query :

‘Select CONVERT( DATE , GETDATE())’

inside the component ‘GET CURRENT DATE’ ,depicted on [Figure 15] .

Figure 15

For instance , if we write this query to SSMS ,[Figure 16], the execution of the query returns the date without including time and also to the format of year-month-day.

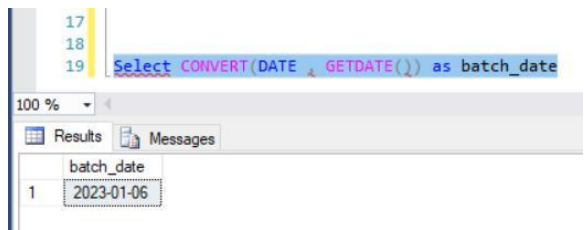


Figure 16

Afterwards, the result of this date is saved to variable named “rvardate”.At the tab variables [Figure 17] ,it is created a new variable named “rvardate” [Figure 18] and the result of the date with the format "YYYY-MM-DD" is going to be stored to this variable [Figure 19] , with the intention to use this variable at the following steps.

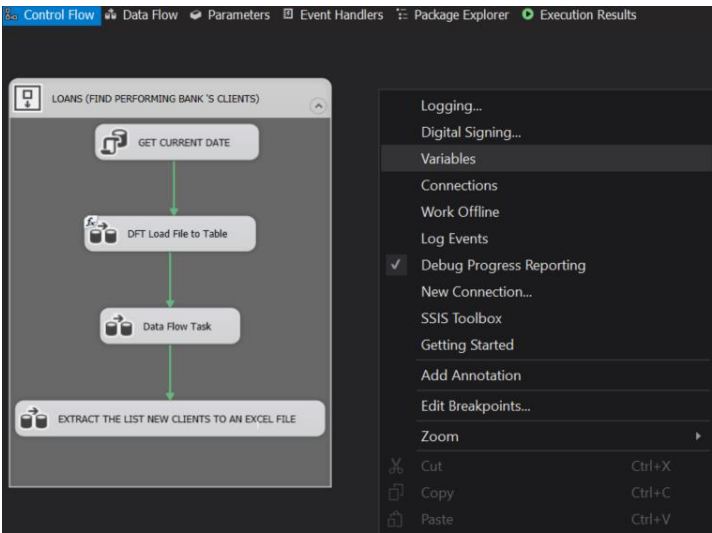


Figure 17



Figure 18

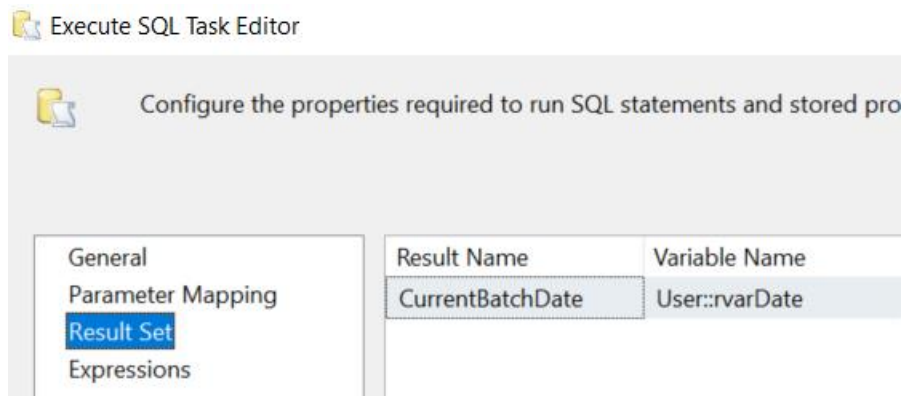


Figure 19

Secondly, at the component “DFT Load File To Table” ([Figure 15]) the step that has to be taken is to upload the dataset about loans. Inside a data flow (Blue component at [Figure 20]) task to the Visual Studio , there is the capability to extract the file that we want .

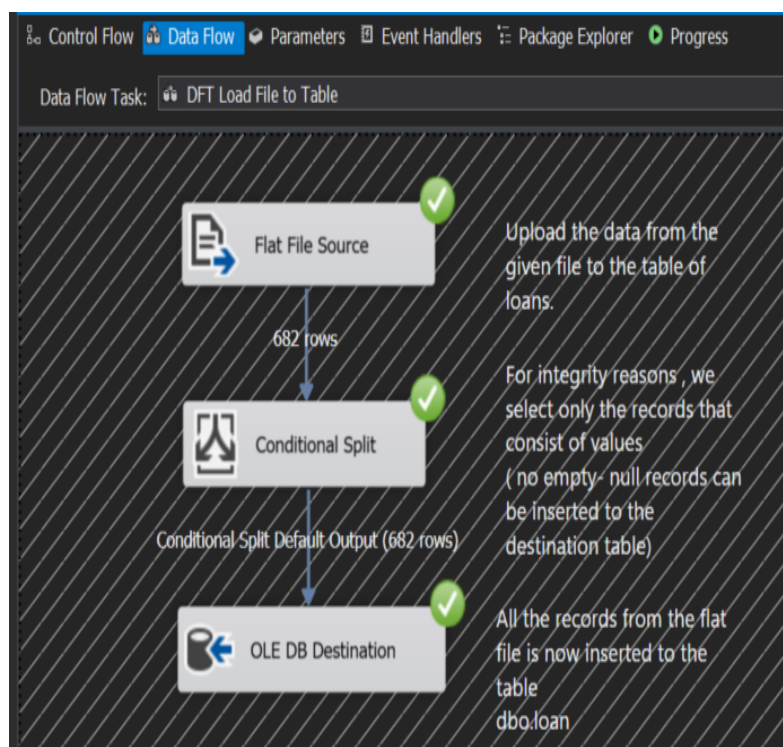


Figure 20

In [Figure 20]) at the Flat File Component, at Component ‘Flat File Source’ the loan file it is uploaded.

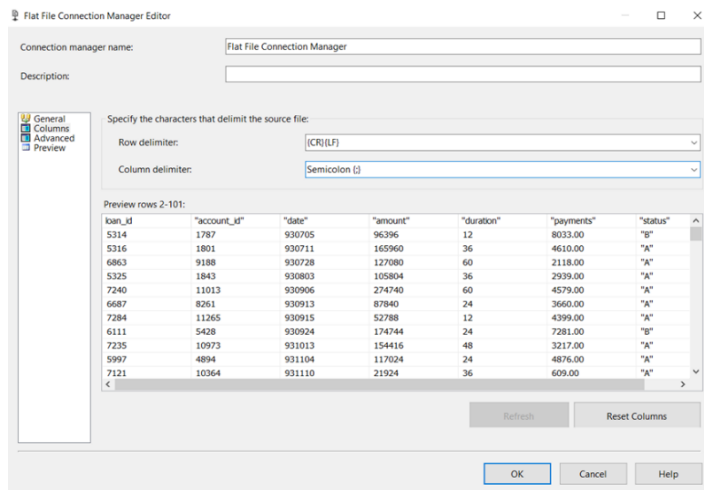


Figure 21

Because the file contains the delimiter “;” between the rows the steps to [Figure 21]) are taken: select the “;” as a column delimiter. Also, in Figure [21], it is clearly depicted how the data is uploaded at the component ‘Flat File Source’.

Furthermore, the following component is ‘Conditional Split’ [Figure 22], that selects only records that have values ( for integrity

purposes ). The expression used is "isnull(status\_id) != false" and it is used to route rows of data to different outputs based on the value of the status\_id column. If case that a record in the column 'status\_id' is NULL, this row data would be sent to the output specified for rows where the expression evaluates to false. In the opposite case where a record in the 'status\_id' is not NULL this row data would be sent to the output specified for rows where the expression evaluates to true. By this way, data is routed to different outputs based on conditions that are specified. As a result, only the rows that contain values are selected.

At the next component named ‘OLE DB Destination’ [Figure 20] all the rows are uploaded from Berka’s dataset file ‘Loans’ that have existing values.

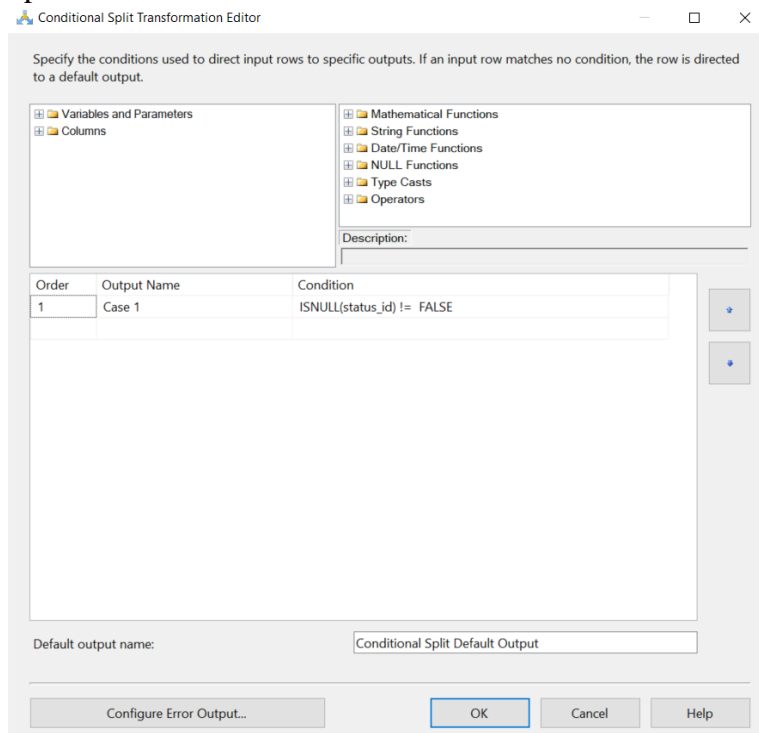


Figure 22

Furthermore, inside the component 'Data Flow Task' [Figure 17], is being followed the procedure depicted on the [Figure 23] :

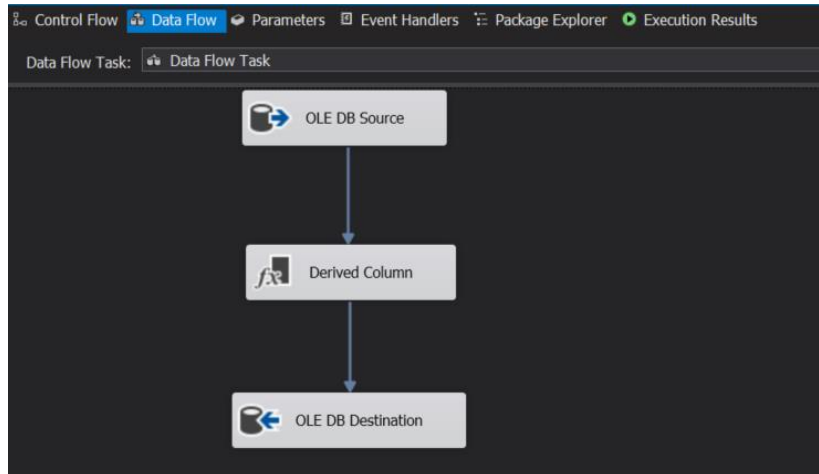


Figure 23

Before the upload of records to the target table `dbo.performing_clients` ( that occurs to the last component 'OLE DB Destination' , only the clients that they are not included to `dbo.performing_clients` have to be selected from the table of loans ( table `dbo.loan` ).Aiming to serve this need, a following SQL code written inside the component of 'OLE DB Source' ( Figure 24 and Table 1 ) :

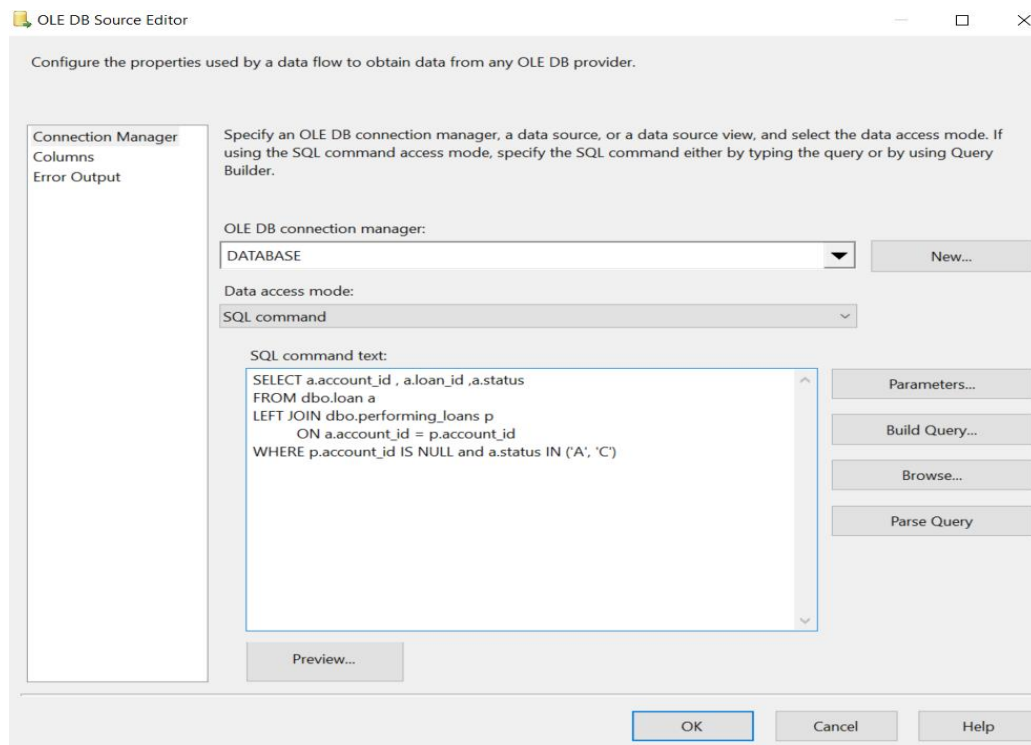


Figure 24

```
SELECT a.account_id , a.loan_id ,a.status
FROM dbo.loan a
LEFT JOIN dbo.performing_loans p
    ON a.account_id = p.account_id
WHERE p.account_id IS NULL and a.status IN ('A', 'C');
```

Table 1: T-SQL code

Using this particular T-SQL query , it managed to obtain-select the records of all clients that their status in 'A' or 'C' and also, they are not included (yet ) to the target table of performing clients . The use of 'LEFT JOIN' aims to sending back all the rows from the table dbo.loan , and any matching rows from the table dbo.performing\_loans. 'WHERE' is utilized for filtering the results to exclusive include rows where there is no matching "account\_id" in the "dbo.performing\_loans" table, and the "status" in the "dbo.loan" table is either 'A' or 'C'.All in all , this T-SQL query is essentially selecting all rows from the dbo.loan table where there is no corresponding row in the "dbo.performing\_loans" table with a matching "account\_id", and the "status" is either 'A' or 'C'

At next component 'Derived Column' depicted on the [Figure 23] , it is added the batch date from our variables [ Figure 25].

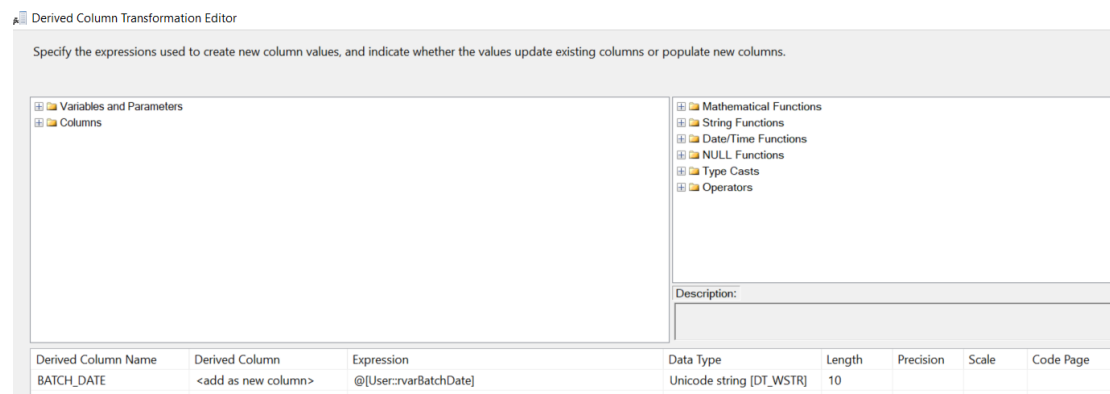


Figure 25

The information of batch date is not taken from our dataset , so that is the reason that we found it and at this step it is added , with the intention to be uploaded to our target table .That happens to the component ‘OLE DB DESTINATION’.

In the final step our procedure that is the component ‘ EXTRACT THE LIST NEW CLIENRS TO AN EXCEL FILE’ [ Figure 17] , it is created a selection of the records that are inserted to the table dbo.performing\_loans and the day of their insertion to table is the current date .Particularly , we upload to an empty csv file all these records , since the purpose of the business intelligence requirement is to have( every time that the table loan is updated ) a list of the clients that their payment performance is good in order to inform them and provide them bank products that they may interested in purchasing . The important improvement is that the clients are informed only one time. The list of the clients imported to the csv does not contain any client that has been already informed. Sending too many marketing messages may violate laws and regulations regarding unsolicited communication, so in this way it is avoided to inform bank’s client more than one time with marketing message. Moreover, the csv file that it is created every time that the package in the Visual Studio is executed is named with the title ‘Updated\_list\_of\_Performing\_clients\_00000000’ where the zero will have the date of the execution. So, if the date of package execution is 3 of January 2023, the file will be named like. ‘Updated\_list\_of\_Performing\_clients\_20230103’. The reason is because the name of the csv will give the business users easily the information of the date that each csv is created.

This is created by selecting the properties option of the Connection Manager of the CSV and creating the ‘Properties’ option [ Figure 26] and adding the variables of our project in a way that will be updated while the package is executing and the name of the CSV will include the current day of the execution of the package [Figure 27]

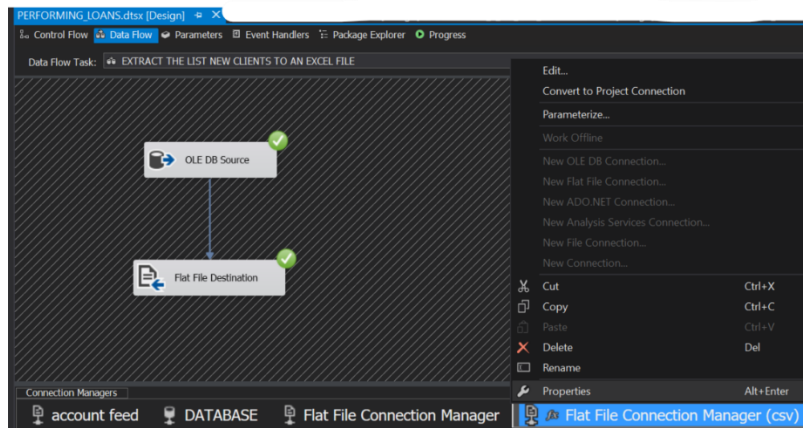


Figure 26

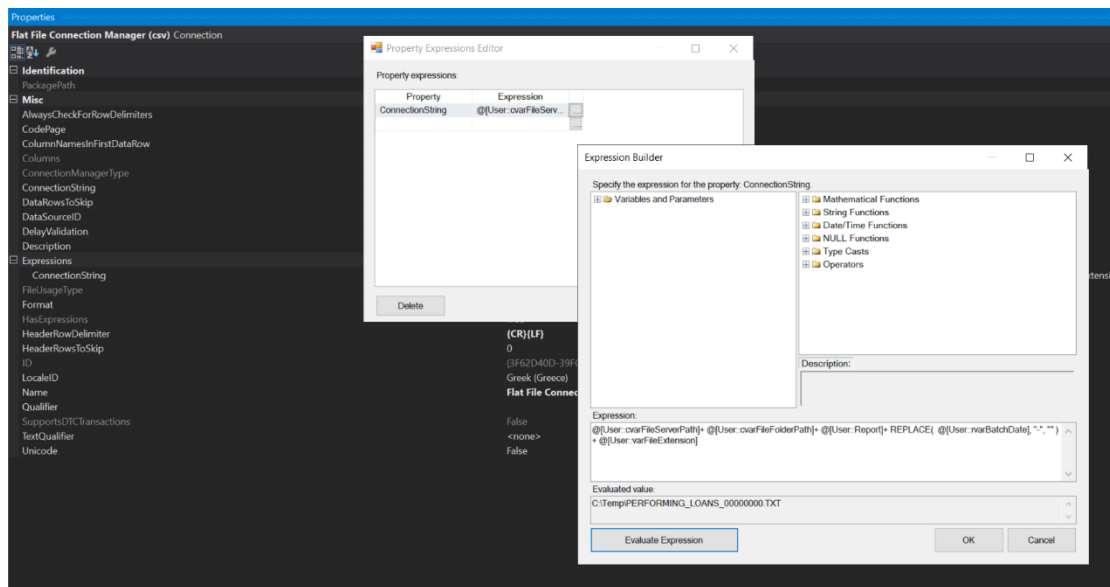


Figure 27

All in all , the previous procedure is the solution of the Business Intelligence problem analyzed above. Every single time that this package run to Visual Studio, there will be created a list with new ‘performing’ clients who they will be informed about the new marketing programs.

## 7.7 DISCUSSION OF RESULTS

It worth mentioning that customer segmentation is a common business intelligence problem that is often solved with data warehousing, which constitutes a powerful approach for businesses to divide a company's customer base into smaller groups based on specific characteristics. By classifying customers based on behaviors such as spending and income, businesses are able to tailor their products, services, and

marketing efforts more effectively retain customers and increase profits. This approach is considered more effective than others because it takes into account the unique behaviors of different customer segments.

The conclusions and results of a customer segmentation constitute an important enhancement on the performance of businesses, in terms of Business Intelligence, for many reasons but mainly because it gives the opportunity to businesses to deeper comprehension of their customers and tailor and promote efficiently their products, services, and marketing efforts to specific groups of customers [54].

At this BI exercise , implemented with Data Warehouse techniques and tools concerns the need of having , every single time that it is needed , a new list with the ‘good’ clients be informed by new Bank products. It created a table that consists of all this information from good clients. Also, the efficiency is to not store repeatedly the same clients every time that this table is going to be updated only with new records. In this way, we keep track of all ‘good clients’ in case we need them again in the future for some different operation and need, we preserve storage on the server Storage by not having the same clients in more than one record. Moreover, every time that this SSIS package is executed , besides updating the target table that keep track of “good” customers , it will be created an excel file , which name will contain the current date of package’s execution date . This excel file will contain only the new “good” clients that they will be inserted to final table .Afterwards , business users that they are interested in this specific list they will be able simply to obtain the information they want .

The data warehousing system used in this thesis was able to collect and store an amount of customer data from the Berka dataset focusing on loan data. The customer segmentation exercise was designed to identify specific groups of customers (segments) that can be targeted with tailored marketing campaigns or that can be used to optimize the ETL process to better handle data from specific segments.

The results of the customer segmentation exercise showed that data warehousing can be used to identify specific customer segments that can be used to improve business intelligence and ETL processes. These segments were characterized by their unique patterns of behavior and characteristics, and it was found that they can be used to target specific groups of customers with tailored marketing campaigns or to optimize the ETL process to better handle data from specific segments.

It is an undeniable fact that Data warehousing can assist businesses to automate the process of customer segmentation by integrating data from various sources and enabling real-time analysis and reporting. Moreover, the use of DW can lead to easy data retrieval and analysis, and also can be easily offered extra help to other aspects of the Data Science ‘umbrella’.

In conclusion, this thesis has highlighted the potential of data warehousing in supporting business intelligence and ETL processes through customer segmentation, making the decision-making procedure more customer centric .

# 8 CONCLUSIONS

Today's world is rapidly changing due to advances in technology, while the world is becoming more globalized and interconnected. Both Industrial and technological growth ushered in the industrial revolution, marked by significant advancements in the industry. This fact has already led to an impressive increase in the amount of data available, which has in turn led to the outcome of creating of new technologies and techniques. Additionally, the rapid pace of change has created new opportunities for businesses to capitalize on, as well as new challenges to overcome.

The field of science and information technology is constantly changing, with new innovations and approaches are continually emerging. The field is continually and constantly developing and changing in an extremely rapid manner. Data warehouses enable businesses to store and analyze a great deal of data from multiple sources, while business intelligence offers insights into customer behavior, market trends, and other relevant information. By leveraging data warehousing for business intelligence cases, businesses can make well-informed decisions and ameliorate their operations.

Data warehouse problems cannot function without ETL processes. These processes enable businesses to acquire data from multiple sources, convert it into a suitable format for analysis, and then store-load it in the Datawarehouse tables. Data warehousing constitutes the process of the efficient retrieval and insightful analysis of data facilitated by organizing and storing data.

In the future, this Business Intelligence exercise can be expanded in several ways. One idea is by incorporating machine learning techniques to further optimize the ETL process and improve the accuracy of customer segmentation. Furthermore, expanding the customer segmentation exercise can be achieved by including more data sources, such as social media and web analytics data, to gain a more comprehensive view of the customer. Also, this thesis could be expanded to include case studies from different industries to gather more insights on how data warehousing is being used in different settings. Moreover, one more idea could be the investigation the use of customer segments in personalized marketing campaigns, and the evaluation the effectiveness of these campaigns in terms of customer engagement and conversion rates.

All in all, data warehousing and business intelligence are essential tools for businesses today.

# 9 REFERENCES

- [1] D. Bentley, "Business Intelligence and Analytics," 2017.
- [2] A. Gupta, "Business Intelligence: Concepts, Components Tools, Techniques, Benefits and Challenges," 2017.
- [3] H. M. Botos, "Business Intelligence and Competitive Intelligence: The Evolution of The Terms," *Research and Science Today*, vol. 16, pp. 56-62, 2018.
- [4] H. P. Luhn, "A business intelligence system," *IBM Journal of research and development*, vol. 2, pp. 314-319, 1958.
- [5] B. Wieder and M.-L. Ossimitz, "The impact of business intelligence on the quality of decision making – a mediation model," *Procedia Computer Science*, vol. 64, pp. 1163-1171, 2015.
- [6] R. Tunowski, "Business intelligence in organization. Benefits, risks and developments," vol. 16, 2015.
- [7] S. Popescu, "Business Intelligence Solutions - a Way of General Improvement of Efficiency and Effectiveness," *REVISTA DE MANAGEMENT COMPARAT INTERNATIONAL/REVIEW OF INTERNATIONAL COMPARATIVE MANAGEMENT*, vol. 13, no. 1, pp. 88-95, 2012.
- [8] G. Gangadharan and S. Swami, "Business intelligence systems: design and implementation strategies," *26th International Conference on Information Technology Interfaces*, vol. 1, pp. 139-144, 2004.
- [9] J. O. Chinyere, "Data warehouse: A tool for organizational efficiency," *Nigerian Journal of Management Sciences*, vol. 23, no. 2, 2022.
- [10] K. Jagadeesh, "Business Intelligence Tools for Process Optimization in a Public Electricity Company," 2020.
- [11] J. O. CHINYERE, "DATA WAREHOUSE: A TOOL FOR ORGANIZATIONAL EFFICIENCY," vol. 23, 2022.
- [12] M. a. M. G. a. o. Velicanu, "Database Vs Data Warehouse," *Revista Informatica Economic*, vol. 3, p. 43, 2007.
- [13] M. Velicanu and G. Matei, "Database Vs Data Warehouse," *Revista Informatica Economic*, vol. 3, p. 43, 2007.

- [14] D. Cardon, "Database vs data warehouse: A comparative review," *Health catalyst*, 2014.
- [15] J. A. R. Rodero, M. Piattini and J. A. T. Alvarez, "The Audit of the Data Warehouse Framework.," in *DMDW*, 1999, p. 14.
- [16] W. Inmon, "Security in the Data Warehouse/Internet Environment," *IS Audit and Control Journal*, pp. 8-11, 1997.
- [17] C. Ballard, D. Herreman, D. Schau, R. Bell, E. Kim and A. Valencic, Data Modeling Techniques for Data Warehousing, IBM Corporation International Technical Support Organization San Jose, 1998.
- [18] P. Geetidarshini and B. Sitanatha , "Data Warehouse and Data Modeling Techniques," *International Journal of Research in Engineering and Science (IJRES)*, vol. 6, no. 8, pp. 2320-9356, 2018.
- [19] K. Rajasekaran, P. Saravanan and P. Prabakaran, "Comparative study on data warehouse tables and schema - An Overview," *Adalya*, vol. 9, pp. 126-132, 2020.
- [20] A. Abello, J. Samos and F. Saltor, "A Framework for the Classification and Description of Multidimensional Data Models," in *Database and Expert Systems Applications*, Berlin, Heidelberg, Springer Berlin Heidelberg, 2001, pp. 668--677.
- [21] D. Anindya and T. Helen, "The cube data model: a conceptual model and algebra for on-line analytical processing in data warehouses," *Decision Support Systems*, vol. 27, no. 3, pp. 289-301, 1999.
- [22] K. Amandeep, "Data Warehousing, Data Mining, OLAP and OLTP Technologies Are Indispensable Elements to Support Decision-Making Process in Industrial World," *International Journal of Scientific and Research Publications*, vol. 5, no. 5, 2015.
- [23] Z. Nebić and M. Viljan , "Data warehouse for an e-learning platform," *Volume II*, pp. 415-420, 2010.
- [24] P. Kaur and P. Kaur, "New Approach of Computing Data Cubes in Data Warehousing," *International Journal of Information & Computation Technology. ISSN*.
- [25] P. Zhao, X. Li, D. Xin and J. Han, "Graph cube: on warehousing and OLAP multidimensional networks," in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, 2011, pp. 853-864.
- [26] C. Z. a. R. M. a. V. V. a. M. D. M. Radulescu, "A multidimensional data model and OLAP analysis for soil physical characteristics," vol. 5109, pp. 25--29, 2008.
- [27] S. Chaudhuri and U. Dayal, "An overview of data warehousing and OLAP technology," *ACM Sigmod record*, vol. 26, pp. 65-74, 1997.
- [28] S. S. a. S. A. Venkatraman, "Formalization of an IS-A based extended nested relation data model," *Information Systems*, vol. 20, 1995.

- [29] D. Calvanese, M. Lenzerini and D. Nardi, "Description logics for conceptual data modeling," *Logics for databases and information systems*, pp. 229-263, 1998.
- [30] D. Dumitriu and M. A.-M. Popescu, "Enterprise architecture framework design in IT management," *Procedia Manufacturing*, vol. 46, pp. 932-940, 2020.
- [31] T. a. P. M. a. K. K. a. D. P. a. D. P. SHARMA, "An introduction to data warehousing and OLAP: pros & cons," *Eureka*, vol. 2581, 2018.
- [32] K. Lyko, M. Nitzschke and A.-C. Ngonga Ngomo, "Big data acquisition," *New horizons for a data-driven economy*, pp. 39-61, 2016.
- [33] G. Kumar, S. Basri, A. A. Imam, S. A. Khowaja, L. F. Capretz and A. O. Balogun, "Data Harmonization for Heterogeneous Datasets: A Systematic Literature Review," *Applied Sciences*, vol. 11, p. 8275, 2021.
- [34] S. Mubeen, S. A. Asadollah, A. V. Papadopoulos, M. Ashjaei, H. Pei-Breivold and M. Behnam, "Management of service level agreements for cloud services in IoT: A systematic mapping study," *IEEE access*, vol. 6, pp. 30184-30207, 2017.
- [35] T. Phungtua-Eng and S. Chittayasothorn, "Slowly Changing Dimension Handling in Data Warehouses Using Temporal Database Features," *Asian Conference on Intelligent Information and Database Systems*, pp. 675-687, 2019.
- [36] T. Manjunath, R. S. Hegadi, I. Umesh and G. Ravikumar, "Realistic analysis of data warehousing and data mining application in education domain," *International Journal of Machine learning and computing*, vol. 2, p. 419, 2012.
- [37] G. Melnik, K. Read and F. Maurer, "Suitability of fit user acceptance tests for specifying functional requirements: Developer perspective," *Conference on Extreme Programming and Agile Methods*, pp. 60-72, 2004.
- [38] A. Abran, J. W. Moore, P. Bourque, R. Dupuis and L. Tripp, "Software engineering body of knowledge," *IEEE Computer Society, Angela Burgess*, p. 25, 2004.
- [39] A. R. Al Essa and C. Bach, "Data Mining and Warehousing," *American Society for Engineering Education (ASEE Zone 1) Journal*, 2014.
- [40] A. Kotopoulos, "Best practices for real-time data warehousing," *Oracle Corporation, Redwood Shores (USA) 2012*, 2012.
- [41] S. H Ali El-Sappagh, A. M Ahmed Hendawi and A. Hamed El Bastawissy, "A proposed model for data warehouse ETL processes," 2017.
- [42] F. S. Esmail, "A Survey of Real-Time Data Warehouse and ETL," *Management*, vol. 9, pp. 3-9, 2014.
- [43] M. Souibgui, F. Atigui, S. Zammali, S. Cherfi and S. B. Yahia, "Data quality in ETL process: A preliminary study," *Procedia Computer Science*, vol. 159, pp. 676-687, 2019.

- [44] J. Shim, P. Scheuermann and R. Vingralek, "Dynamic caching of query results for decision support systems," *Proceedings. Eleventh International Conference on Scientific and Statistical Database Management*, pp. 254-263, 1999.
- [45] V. Gour, S. Sarangdevot, G. S. Tanwar and A. Sharma, "Improve performance of extract, transform and load (ETL) in data warehouse," *International Journal on Computer Science and Engineering*, vol. 2, pp. 786-789, 2010.
- [46] "Credit Card Analysis of Czech Bank," 11 2002. [Online]. Available: <https://webpages.charlotte.edu/mirsad/itcs6265/group1/index.html>.
- [47] M. B. Shahid, U. Sheikh, B. Raza and Q. Javaid, "Application of data warehouse in real life: State-of-the-art survey from user preferences' perspective," *International Journal of Advanced Computer Science and Applications*, vol. 7, 2016.
- [48] S. Amann, S. Proksch, S. Nadi and M. Mezini, "A study of visual studio usage in practice," in *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, 2016, pp. 124-134.
- [49] M. V. Vladimir Šimović, "Applying SQL database query to access SQL server 2019--visual studio 2019," 2019.
- [50] D. a. S. K. a. Y. S. Kabakchieva, "Latest Trends in Business Intelligence System Development," p. 212, 2013.
- [51] "SQL Server Management Studio," [Online]. Available: [http://ptgmedia.pearsoncmg.com/images/9780672328244/samplechapter/0672328240\\_ch03.pdf](http://ptgmedia.pearsoncmg.com/images/9780672328244/samplechapter/0672328240_ch03.pdf).
- [52] I. C. a. F. J. E. Italiano, "Synchronization options for data warehouse designs," *Computer*, vol. 39, pp. 53--57, 2006.
- [53] B. Cooil, L. Aksoy and T. L. Keiningham, "Approaches to customer segmentation," *Journal of Relationship Marketing*, vol. 6, pp. 9-39, 2008.
- [54] N. Patankar, S. Dixit, A. Bhamare, A. Darpel and R. Raina, "Customer Segmentation Using Machine Learning," 2021.
- [55] R. Tunowski, "Business Intelligence in Organization. Benefits, Risks and Developments," *Przedsiębiorczość i Zarządzanie*, vol. 16.
- [56] S. U. Rehman, "A Visual Studio Add-In for Software Component Services in Smart Devices," 2011.
- [57] B. M. Ramageri, "Data mining techniques and applications," *Indian Journal of Computer Science and Engineering*, vol. 1, 2010.
- [58] C. Z. Radulescu, M. Radulescu, V. Vlad and D. M. Motelica, "A multidimensional data model and OLAP analysis for soil physical characteristics," *proceeding of: 9-th WSEAS*

*International Conference on Mathematics & Computers in Business and Economics (MCBE'08)*, vol. 5109, pp. 25-29, 2008.

- [59] R. Mistry and S. Misner, *Introducing Microsoft SQL Server 2014*, Microsoft Press, 2014.
- [60] H. Homayouni, S. Ghosh and I. Ray, "An approach for testing the extract-transform-load process in data warehouse systems," *Proceedings of the 22nd International Database Engineering & Applications Symposium*, pp. 236-245, 2018.
- [61] P. Castro, M. B. Alves and I. Rostomashvili, "Key performance indicators in a high educational institution : a business intelligence approach," 25 11 2010. [Online]. Available: <http://hdl.handle.net/20.500.11960/1549>.
- [62] S. a. P. S. a. N. S. a. M. M. Amann, "A study of visual studio usage in practice," vol. 1, pp. 124--134, 2016.
- [63] M. a. M. G. Velicanu, "Database versus Data Warehouse," *Editura Economică*, 2007.
- [64] P. a. L. X. a. X. D. a. H. J. Zhao, "Graph cube: on warehousing and OLAP multidimensional networks," pp. 853--864}, 2011.