# Fundamentals

## Data Warehousing and Business Intelligence

### Data Warehousing

In this chapter I will give several definitions of Data Warehousing and its application and need. Furthermore, I will provide a brief overview about the current trends in Data Warehousing.

Bill Inmon, the ‘father of Data Warehousing’ defines a data warehouse as ‘a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management’s decision making process’ (Inmon, 2005).

Another famous definition is given by Ralph Kimball who defines a Data Warehouse as ‘a system that extracts, cleans, conforms, and delivers source data into a dimensional data store and then supports and implements querying and analysis for the purpose of decision making’ (Kimball & Ross, 2013).

Vincent Rainardi defines a Data Warehouse as ‘a System that *retrieves* and *consolidates* data *periodically* from the source systems into a *dimensional or normalized data store*. It usually keeps years of *history* and is *queried* for *business intelligence* or other *analytical* activities. It is typically updated in batches, not every time a transaction happens in the source system’ (Rainardi, 2011). In this chapter we will take a closer look on this definition to get an idea of how a Data Warehouse is used.

The retrieval of the data stored in a Data Warehouse is often described as Extract – Transform – Load (ETL). In the first step data is extracted from one or several source systems. After that the data is transformed using data quality rules which are defined by the developer. This might be the handling of missing data or removal of duplicates. In the last step of ETL retrieval the data is loaded and stored into the target system (Rainardi, 2011).

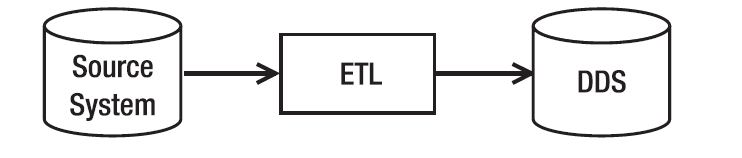


Figure : ETL (Rainardi, 2011)

Data Consolidation means harmonizing data from different data sources. For example, the time columns from different sources can be in different formats or data is displayed in various unity of measure or it contains the same information but has different names. To use all the different kinds of data sources for analysis data consolidation is very important (Rainardi, 2011).

Usually, the retrieval and consolidation of data happen at regular intervals e.g. daily or monthly depending on the source system’s update frequency as well as the business need for analysis. If reports and statistics for the data is only needed once a month it is not required to update the data daily. This can help to improve performance (Rainardi, 2011).

A Data Warehouse usually uses the concept of Dimensional Data Storage. A dimensional data store is denormalized and consist of a group of fact tables connected to so-called dimension tables. A widely used database schema in dimensional data stores is a star schema or a snowflake schema (Rainardi, 2011). Figure 2 shows such a star schema. In its center there is an example fact table sales order item. Fact tables contain measurements and metrics which are displayed mostly in numeric values. They also contain foreign keys which connect the fact tables with the corresponding dimension tables. Dimension tables contain descriptive information about the data (Kimball & Ross, 2013).

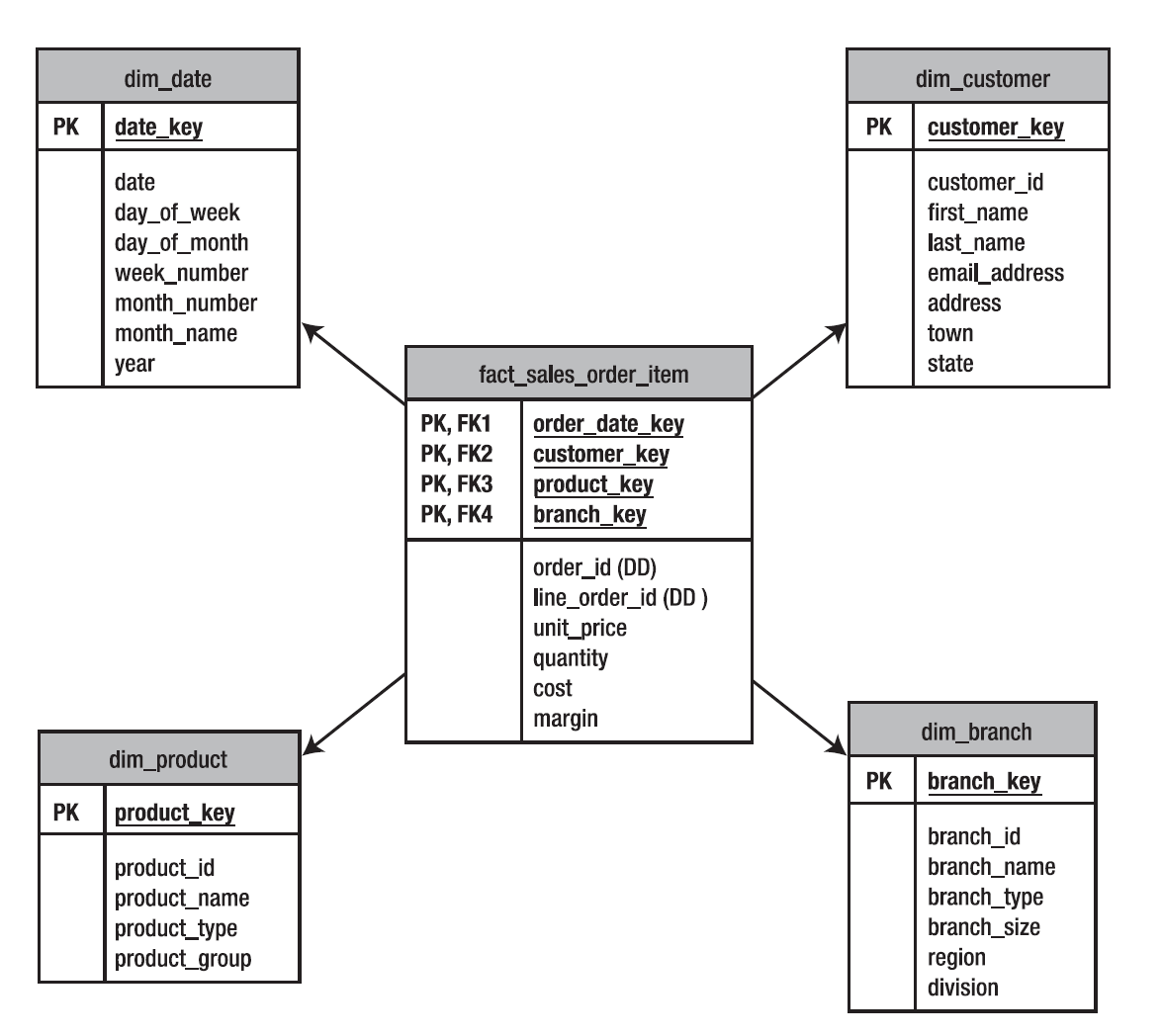


Figure : Star Schema (Rainardi, 2011)

Another key characteristic of a Data Warehouse is the capability to store large amounts of historical data whereas most source systems do not store long-term transactional data because it is not necessary for their functionality. This is why Data Warehouses need large storage capacities as well as technical database features such as table portioning, parallel query or in-memory databases. A challenge that comes up with the storage of historical transactional data is that master data that describes transactional data is changing over time. Unlike most transactional systems, data warehouses also store master data using a technique called slowly changing dimensions (SCD) where historical information about dimensional data is stored either in rows or columns (Rainardi, 2011).

A Data Warehouse also includes the ability to query data more effective than a source system. It is because transactional system has its focus on performing several database transactions such as update, insert, delete or select while Data Warehouses can perform complex queries in a very fast way due to a denormalized database structure. Additionally, in a Data Warehouse you have stored all the data from various source systems so you can be sure to access not only partial data (Rainardi, 2011).

### Business Intelligence

Data which is stored in a Data Warehouse is typically used for analysis. The collection of those activities is called Business Intelligence (BI). Forrester Research defines Business Intelligence as “a set of methodologies, processes, architectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making” (Forrester Research, 2008). This includes the preparation and storage of data using a data warehouse as well as the data usage.

The data usage can be grouped into three different categories:

* Reporting
* Online Analytical Processing
* Data Mining

Reporting describes the creation of key performance indicators and other business relevant figures and the deployment as a static report.

Online Analytical Processing (OLAP) is the activity of analyzing data interactively in order to make business decisions. To perform OLAP you make use of the technical concept of a multidimensional database structure. You can think of a multidimensional structure as a cube with each cell representing aggregated data related to elements along each of its dimensions (O'Brien & Marakas, 2011; Rainardi, 2011). Figure shows a model of an OLAP cube with the three dimensions customer, store and time. If a user wants to analyze the profit on day X, customer segment Y and store Z the OLAP processor accesses the associated cell of the cube. OLAP also comes with the ability to perform Slice and Dice operations (see Figure 4). Slicing means selecting a value on one of the dimensions for example a specific date and taking of a slice from the data cube. Dicing describes the creation of smaller cubes within the OLAP cube by selecting certain value ranges for dimensions. Further operations are drill-down which means zooming into the cube according to dimension hierarchies or pivoting.

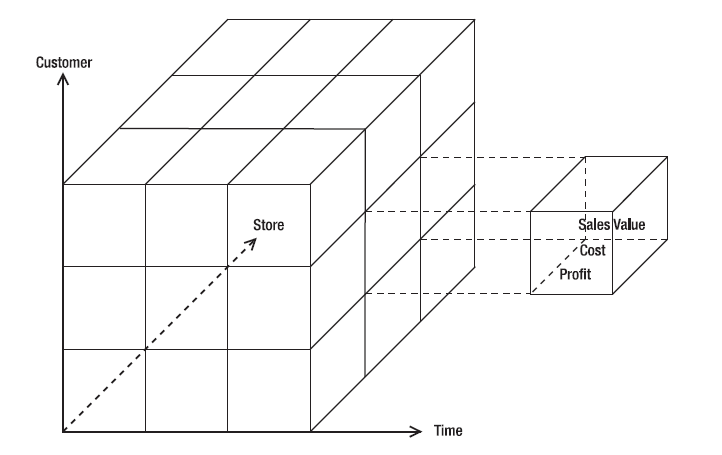
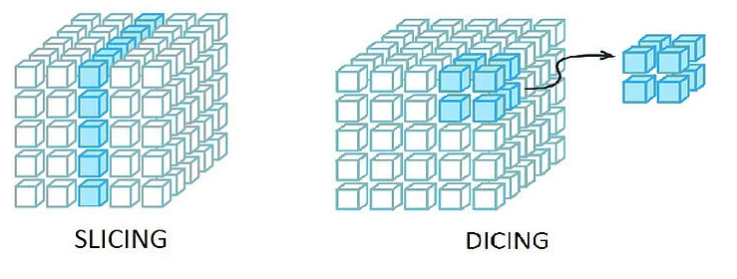


Figure : OLAP Cube (Rainardi, 2011)



Data Mining which is the third type of application in BI is described in chapter 2.3

### Cloud Business Intelligence

Cloud Computing is one of the biggest trends in IT in the last years. It provides the capability for companies to scale their infrastructure dynamically from third party vendors. In an expanding economy of scale, virtualized and dynamically-scaling computing power, storage capacity, platforms and services on demand have become necessary to survive (Thompson & van der Walt, 2010). The National Institute of Standards and Technology lists three cloud computing business models:

* Infrastructure as a Service (IaaS) such as virtual machines, servers or networks
* Platform as a Service (PaaS) such as web-environments for developers or databases
* Software as a Service (SaaS) such as e-mail, communication tools or games (Mell & Grance, 2011)

As more and more companies move their enterprise infrastructure into the cloud, many vendors of BI software introduced BI and Data Warehouse as cloud services. Since implementing and maintaining BI is often very expensive, cloud computing can offer a solution to reduce coasts. As the data staging layer is often the most challenging one in terms of performance and maintaining, this layer is very likely to be transformed into a cloud service where different data sources will be extracted and cleansed. The ability of clouds to be dynamically scalable is very useful for this scenario It is also very useful to outsource the access layer that includes the reporting and analytics activities since business users can access them easily though a portal including mobile access. Additionally, appropriate data governance can be ensured (ElMalah & Nasr, 2019). Especially for Data Mining Tasks, business require fast and powerful solutions to perform analytics or predictions.

## Data Mining

Due to the increasing digitalization and use of computers in various fields, the amount of data generated by information systems and sensors is also increasing. In the past, a lot of work has been done to effectively collect and store this data. However, only a small portion of this amount of data is used, leaving much unused, potentially valuable information unnoticed. The reasons for this are, on the one hand, the large amount of data itself and, on the other hand, the data structures whose primary function is efficient data storage and which were not created for the purpose of analysis. However, the need for knowledge and information that can be generated from company and customer data is growing. In today's business world, the possession of this knowledge can often already represent a competitive advantage over competitors (Kantardzic, 2011).

The process of finding new, valuable and non-trivial information in large amounts of data is known in science as data mining. It combines the expert knowledge of people and their ability to describe and analyze facts, as well as the technical possibilities offered by today's computers to apply this to large amounts of data. The primary goal of these methods is to find data patterns that can be understood and interpreted by humans in order to gain knowledge from them (Kantardzic, 2011).

The applications of data mining can be divided into six different categories, these are:

* Classification (learning a model using existing data sets, which is able to classify new data into one of several defined classes)
* Regression (teaching a model that is capable of predicting numerical values for data)
* Clustering (division of data into logical categories that did not exist before)
* Summarization (finding a compact description for a set of data)
* Dependency modeling (building a model that finds dependencies between variables or between attribute values within data)
* Change and Deviation detection (detection of outliers and significant changes within the data) (Kantardzic, 2011)

In the further course of this work, the focus will be on the application field of classification, since the problem described at the beginning is a classification problem.

Data Mining problems can also be grouped into supervised learning, unsupervised learning, semi-supervised-learning and active learning. Supervised learning is synonymous with classification and means learning through labeled data in the training set. Unsupervised learning describes the clustering of data. It is called unsupervised because the input data do not have labels. Usually clustering is used to find classes within data. Semi-supervised learning takes both labeled and unlabeled data as the basis for learning the model. This approach is often chosen when a classifier is trained with the labeled data and the boundary between the classes is determined with the unlabeled examples. Finally, active learning is an approach in which users actively contribute to the learning process. For example, domain experts are asked to label data sets. The goal is to increase model quality through the active assistance of human users.

### Classification

Classification by means of data mining takes place in many areas of life. Examples are fraud detection, risk analysis for banks or medical diagnoses. Han defines classification as a "form of data analysis that extract models describing important data classes" (Han, 2012) These models, which are also called classifiers, serve the purpose of determining categorical attributes, so-called labels. These attributes can take on discrete values such as 1,2 or 3, but do not represent ordinal quantities with a fixed order, but can be used synonymously for a state A, B or C. The prediction of labels with numerical values, on the other hand, is usually solved by using statistical methods of regression (Han, 2012; Kantardzic, 2011).

The process of classification has usually 2 phases. It consists of a learning and a classification process. This is exemplified in Figure . Based on a training dataset, a classification algorithm is trained and a model is created. This training set consists of a subset of the entire data set. A single expression of the attributes and the corresponding class, corresponding to a row of the data set, is called a tuple in the following.

Reference List

ElMalah, K., & Nasr, M. (2019). Cloud Business Intelligence. *International Journal of Advanced Networking Applications, 10*(06), 4120–4124.

Forrester Research (2008). *Topic Overview: Business Intelligence.* Retrieved November 17, 2021, from https://www.forrester.com/report/Topic-Overview-Business-Intelligence/RES39218.

Han, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). *Morgan Kaufmann series in data management systems.* Waltham, MA: Morgan Kaufmann/Elsevier.

Inmon, W. H. (2005). *Building the data warehouse* (Fourth edition). *Wiley technology publishing Timely, practical, reliable.* Indianapolis, Indiana: Wiley.

Kantardzic, M. (2011). *Data Mining.* Hoboken, NJ, USA: John Wiley & Sons, Inc.

Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling* (3rd Edition). New York, NY: John Wiley & Sons.

Mell, P. M., & Grance, T. (2011). *The NIST definition of cloud computing.* Gaithersburg, MD.

O'Brien, J. A., & Marakas, G. M. (2011). *Management information systems* (10th ed.). New York: McGraw-Hill Irwin.

Rainardi, V. (2011). *Building a Data Warehouse: With Examples in SQL Server. Books for professionals by professionals.* New York, NY.: Apress Springer.

Thompson, W. J., & van der Walt, J. S. (2010). Business intelligence in the cloud. *SA Journal of Information Management, 12*(1).