## The Business Intelligence Concept

(Connolly & Begg, 2015) defines Business Intelligence as *“an umbrella term that refers to the processes for collecting and analyzing data, the technologies used in these processes, and the information obtained from these processes with the purpose of facilitating corporate decision making”.*  The term Business Intelligence is a relatively new one as it appeared around mid - 90s. However, systems that provide Business Intelligence exist since early 1980s – they used to be called Enterprise Information systems (EIS) (Sharda, Delen, Turban, & King, 2015). EIS were the first systems that had the capability of providing advanced data analytics such as forecasting, prediction and ad-hoc reporting. EIS replaced the Management Reporting Systems (MRS) that showed up in the 1960s and had very limited capabilities. As the EIS continued evolving, they were enhanced with additional features such as artificial intelligence and data mining. These systems are now called Business Intelligence systems (BIS) and are a very important element of a modern business in terms of supporting decision making.

## Business Intelligence Systems

(Sharda et al., 2015) defines a Business Intelligence System as system that has four components. A Data Warehouse, Business Analytics, a User Interface and Business Performance Management (BPM). The data warehouse is the central repository of data that have been collected from different internal or external sources. Business Analytics refer to tools that are used for querying, mining or analyzing data from the data warehouse such as OLAP tools or Data Mining tools. The User Interface refers to the dashboards or reports that can be produced by modern reporting tools. Finally the business performance management (BPM) is tightly connected with Business Intelligence as the latter can help BPM to identify strategy requirements, drive performance and monitor achievement. (Fidler, 2016). This project will focus on the data warehouse component, and more specifically of the methods that are used to extract data from data sources and the load the data into the data warehouse, after performing any necessary transformations.

## The Data Warehouse concept

A data warehouse is the main component of a business intelligence system. It’s another database, technically similar to the operational databases, but serves a completely different purpose. While an operational database is optimized to process transactions efficiently and assist the knowledge workers to their day to day tasks, a data warehouse is focused on analytical processes and has the goal to improve decision making. A data warehouse is a decision support system that provides clean and credible data. Businesses are using operational systems to get data in and data warehouses to get data out. (Kimball & Ross, 2013). There are many data warehouse design approaches but two of them are the most prevalent. Bill Inmon, known as the “father of the data warehouse” suggests a centralized data warehouse architecture with dependent data marts while Ralph Kimball advocates the data mart bus architecture with linked dimensional data marts. (paper-43). (figures appendix). The data warehouse is built in the 3d normal form, after an extensive process where the business requirements across all departments of the enterprise are defined. Then data marts that support business processes or specific subjects/functions of the enterprise are built on top of the enterprise data warehouse. On the contrary, Kimball’s architecture follows a bottom-up approach and a different philosophy. The data warehouse is built step by step by creating data marts. Each of the data marts supports a different business process and is modelled by a specific data modelling technique called dimensional modelling. (see appendices). The data marts are linked via conformed dimensions and form the data warehouse. The design approach of the data warehouse affects also the design of the ETL process, as the loading stage is dependent on the underlying data model. The designers of the ETL process should be aware of the architecture of the data warehouse and adjust the design of the loading stage accordingly.

## Extract, Transform, Load (ETL)

ETL stands for Extract, Transform and Load and it’s a term widely used in data warehousing. The ETL system is the backbone of a data warehouse as it is responsible for the extraction of the data from external sources, the cleaning of the data and the loading of the data into the data warehouse. Behrend and Jörg research (as cited in paper 32) found that seventy percent (70%) of the design, implementation and maintenance of a data warehouse is allocated to the ETL system, which is a complex project divided into many subtasks. There are many different methodologies, tools and technologies for ETL development and implementation. ETL design is a significant part of the Business Intelligence lifecycle (Moss and Atre Shaku, 2003). Before implementing any data pipelines , The ETL Team should take into consideration the business requirements and plan the ETL solution accordingly.

The extract step of the ETL process should be planned based on the business needs. The business requirements set by end users define the data sources or specific entities and attributes of an operational system that need to be considered for integration into the data warehouse. Data sources or other database objects that are not useful for analysis should not be considered. The Transform step is also driven by the business needs. Specific business rules are applied at this step. The data are cleaned, conformed and ready to be imported in the data warehouse. The transform step is very important as it is responsible for providing a high level of data quality to the end users. After the data have been transformed, they are loaded to the data warehouse.

The end users want to have easy access to information and they should be able to understand the underlying data model. The Business Intelligence Team should chose a data model that is simple, scalable and efficient and the ETL Team needs then to design the ETL processes to load the data efficiently in the data structure that has been implemented based on the chosen data model.

An additional business requirement that affects all of the steps above is the data latency requirement (Kimball Ralph, 2004). The frequency the data warehouse needs to be updated with fresh data is one of the most important aspects to consider by the design of an ETL solution. The update frequency of a data warehouse / data mart is driven by the business needs and can vary from once per month to real-time. The data latency requirement can have a huge impact in the design of an ETL solution as it can define the ETL Software/Hardware Architecture. Real-time ETL solutions require modern ETL tools that use a different architecture.

## Evolution of ETL Architecture

(Paper 44) identifies three distinct generations of ETL tools. The ETL tools of the first generation were written in the native language of the operating system and did not have their own ETL engine. Most of them were used in mainframes and were written in COBOL or C. These tools were single threaded and didn’t support parallelism. In addition, they required developers with strong programming skills.

The ETL tools of the second generation appeared in the mid -1990s. These tools did not generate code but had their own internal ETL engines. In most cases these tools were using a dedicated ETL server, which acts as a hub server between the data sources and the target databases. The advantage was that the data processing was taking place in the ETL server and not in the source systems. In addition, the developers did not have to know how to programming in different languages (the native language of the operating system) but only in one programming language, the one of the ETL tool. The tools of this generator were supporting limited parallel processing and had a GUI.

The ETL tools of the third generation, are either code generators or engine based but are very powerful and advanced. These tools support multiple data sources and have pre-build connectors for hundreds of applications and APIs. In addition they offer advanced transformations and support specific functions of data warehousing and dimensional modeling such as the automatic creation of a slow changing dimensions, surrogate keys as well as easy deployment and scheduling. These modern ETL tools are able to provide advanced parallel processing and a very user friendly visual environment.

As Business Intelligence evolves and the user community becomes more energetic, the requirement of near real-time ETL has led the ETL vendors to enhance the ETL tools with advances features that support real-time or near real-time ETL. Therefore, in addition to paper 44 work, which was published a decade ago, an emerging fourth generation of ETL systems should be identified. These systems have kept the advanced features of their predecessors but are using advanced techniques for real time extraction, transformation and loading of data, either as standalone products or as add-ons to existing ETL technology. Log based Change Date Capture and Streaming data are some terms that are tightly connected to the fourth generation of ETL tools. These technologies will be extensively described in the following paragraphs.

# Towards real-time Business Intelligence

## First generation of Business Intelligence Systems

In the early days of Business Intelligence Systems, data warehouses were serving a very specific purpose: the storage of historical information across different data sources in a central database. Users would analyze then the historical data with the assistance of Business intelligence applications and take decisions based on their findings. The analysis of the data would answer the question “What happened?” For a local super market this would be translated to “How many products did we sale last week?” or “What was the product with the highest sales on Christmas Eve?”. The update frequency of a data warehouse with fresh data would vary from monthly to weekly basis. Later this would change to once per day, most of the times overnight. The ETL processes could cause a performance impact to the OLTP systems but during the night, the production databases were operating with a low load or they could also go offline if needed. However, users and medium level management had also the need to analyze operational data on a daily basis. Standard reports e.g., total sales of yesterday were generated by SQL queries that were running directly on the operational databases. As this was a bad practice that had a performance impact on the OLTP systems, the Business Intelligence Systems were enhanced with an extra layer: The Operational Data Store (Kimball & Ross, 2013). The Operational Data Store was a staging area in the data warehouse or another database that was hosting a copy of the data sources tables. This database was populated either by ETL tools or by data replication software. Ad-hoc SQL queries or Reporting used as a source the data in the operational data store instead of the OLTP systems. By doing this, the knowledge workers could get the operational data quickly without placing a burden in the production systems. To summarize, the first generation of Business Intelligence Systems had an Operation Data Store layer for operational reporting and a data warehouse that was updated on monthly/weekly/daily basis for historical analysis. As aptly described by the white paper of (GoldenGate, 2009), data warehouses of that generation were *strategic* with emphasis on *reporting.*

## Second generation of Business Intelligence Systems

As more and more businesses are becoming data-driven and the amounts of data are growing at a rapid pace, the business intelligence systems should adapt to the new conditions. The business requirements of the users have changed - the user community demands now not only valid, accurate and cleansed data but also fresh data- a low data latency from the target databases to the data warehouse that allows the decision makers to analyze data quickly and effectively. Many data warehouses as of today belong to this generation. The data warehouse refresh rate has been increased up to hourly basis. In addition, both OLTP and OLAP systems are more powerful than before. This had led to significant changes in the Business Intelligence Architecture – the operational data store is gone and the data warehouses is populated directly from the production systems. Data Warehouse Experts, such as Ralph Kimball (Kimball & Ross, 2013) have removed the operational data store from the architecture they propose. Simple or ad-hoc operational reports use a source either directly the powerful OLTP systems or the data warehouse which is up to date. Dashboards and advanced analytics are now easier and more meaningful to generate, not only because the low data latency but also because the high-end commercial Business Intelligence tools that are now available. GoldenGate (2009) names the data warehouses of this generation as *strategic and predictive*. Business Intelligence is now able to answer the questions “What is happening?” and “What will happen?”

## Third generation of Business Intelligence Systems

The third generation of Business Intelligence Systems consist of the so called Active Data Warehouses. Stephen Brosbt research in 2001 (as cited by (Nguyen Manh Tho & Tjoa Min, 2006))