# Data Extraction, Transformation and Loading (ETL)

According to G. Goos et all., (2003) Data Warehouse (DWs) purpose is to process and facilitate decision making processes. Therefore, the Extraction-Transformation-Loading (ELT) functionality is important to facilitate any query within complex computer systems. ETL processes are liable to extract data from operational data sources, transformation of these data that includes normalization and loading clean data back to DWs. That process is crucial component of DWs data flow as incorrect or misinterpreted data will lead to the inaccurate business decision, therefore data quality must be observed at its early stage of loading. The design of ETL processes structured in 6 tasks:

1. **SOURCE EXTRACTION**

Includes extraction of data from many operational systems. Therefore, to obtain desired dataset more data might be extracted, the size of it depends on actual system and business requirements.

1. **TRANSFORM THE SOURCES**

Includes processes of filtering to required dataset. The filtering tasks will include calculating values originated values, converting between different data formats and codes, sequence numbers automatic generation, elimination of duplicates and eliminating unwanted information. The most complex and time-consuming transformations are Multistage Data Transformation and Pipelined Data Transformation.

1. **JOIN THE SOURCES**

Includes merging different operational sources for unique data loading.

1. **SELECT TARGET TO LOAD**

Includes selection process of the target that must be loaded into the system.

1. **MAP SOURCE ATTRIBUTES TO TARGET ATTRIBUTES**

Includes mapping process of extracted attributes to the corresponding target elements.

1. **DATA LOAD**

Incudes upload of transformed data into DWs.

The vital stage of ELT process is data cleaning with identifying and eliminating errors and data discrepancies in order to increase data quality. According to Goos G, Hartmanis (2003) “Data quality problems are very significant: it has been estimated that poor quality customer data cost U.S. businesses $611 billion a year in postage, printing, and staff overhead.” Therefore the conceptual modeling of ETL processes is beneficial for running and maintaining Data Warehouse.

# Extraction Methods in DWs

The extraction method depends on the source system, warehouse environment and business needs, and this influences the data source, the transportation process and the time need it to update Warehouse. Generally speaking the data for extraction might be badly documented and therefore extraction will be acquired several times. The source system cannot be modified to accommodate needs of the data warehouse extraction processes; therefore two types of extraction identified such as Logical and Physical.

The **Logical extraction** method used when there is no possibility to add additional logic to the source systems to improve its performance. There are two types of logical extraction:

* **Full Extraction** exports currently available data on the source system. An example for a full extraction may be an export file of a distinct table or a remote SQL statement scanning the complete source table.
* **Incremental Extraction** exports data that has been changed since specific time event. This information can be recorded by source system reflecting to the date stamps or by change- capture mechanism. Instead of exporting entire tables it captures and exports only areas with latest implementation. This method is particularly useful on big chunks of data.

The **Physical Extraction** method depends what logical extraction method was applied, also on the source system effectiveness and limitations. There are two mechanisms of physical extraction online from the source system or from offline structure.

* **Online extraction** exports data from the source system or from the intermediate system that stores data in log format. The extracted information must be determined whether it is extracted from the original entity or prepared entity.
* **Offline extraction** exports data outside of original data source system. This data available in redo logs, archive logs or transportable table-spaces or was created by previous extraction routine. There are several structures to be aware of:
  + **Flat files** (contains generic format of outlined data, requires additional information for further managing),
  + **Dump** files (Oracle distinct files, outlined data might be not present),
  + **Redo and archive logs** (information stored in additional dump file),
  + **Transportable table-spaces** (used to extract and transport large volumes of data between Oracle databases), (Oracle, 2014).

**REFERENCES:**

Goos G, Hartmanis, J. and van L.J., 2003. Conceptual Modeling-ER 2003. *Springer-Verlag Berlin Heidelberg*. Available at: http://download.springer.com/static/pdf/54/bok%3A978-3-540-39648-2.pdf?auth66=1421251779\_b8d74ae212dcb392186662a364036b7b&ext=.pdf [Accessed January 14, 2015].

Oracle, 2014. Database Data Warehousing Guide. *Oracle Database Online Documentation*. Available at: http://docs.oracle.com/cd/B19306\_01/server.102/b14223/extract.htm [Accessed January 19, 2015].