**\* ETL Process \***

ETL, which stands for extract, transform and load, is a data integration process that combines data from multiple data sources into a single, consistent data store that is loaded into a [data warehouse](https://www.ibm.com/cloud/learn/data-warehouse) or other target system.

1. **Extract:**

* The first step of the ETL process is extraction.
* In this step, data from various source systems is extracted which can be in various formats like relational databases, flat files etc. into the staging area.
* It is important to extract the data from various source systems and store it into the staging area first and not directly into the data warehouse because the extracted data is in various formats. Hence loading it directly into the data warehouse may damage it and rollback will be much more difficult. Therefore, this is one of the most important steps of ETL process.
* During data extraction, raw data is copied or exported from source locations to a staging area.

1. **Transform:**

* The second step of the ETL process is transformation. In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format.In the staging area, the raw data undergoes data processing. Here, the data is transformed and consolidated. It may involve following processes/tasks:
  + Filtering – loading only certain attributes into the data warehouse.
  + Cleaning – filling up the NULL values with some default values, mapping

e.g., U.S.A, United States, and America into USA, etc.

* + Joining – joining multiple attributes into one. Formatting the data into tables or joined tables to match the schema of the target data warehouse.

e.g.

we join multiple tables from input and load into target system as one single table as per business requirements.

* + Splitting – splitting a single attribute into multiple attributes.
  + Sorting – sorting tuples on the basis of some attribute (generally key-attribute).

1. **Data Cleansing is process under Transform data in required format:**

* While processing flat files to database cleansing is required.
* Data cleansing is also known as data scrubbing. It is the process that set of data correct and accurate.
* **Data cleansing:**

Weeding out unnecessary and unwanted things from incoming data to make it more meaningful and informative.

* **Data Merging:**

Data can be gathered from heterogeneous systems and put together.

* **Data scrubbing:**

Data scrubbing is the process of fixing or eliminating individual pieces of data that are incorrect, incomplete or duplicate before data is passed to end user.

1. **Loading:**

* The third and final step of the ETL process is loading. In this step, the transformed data is finally loaded into the data warehouse.
* Sometimes the data is updated by loading into the data warehouse very frequently and sometimes it is done after longer but regular intervals.
* The rate and period of loading solely depends on the requirements and varies from system to system.
* From here data is take for analysis purpose or for make dashboard purpose to show to client by the business person e.g AWS Quicksight

1. **What is staging area and why are we need in DWH?**

* If target and source database are different and target table volume is high it contains some millions of records in this scenario without staging table, we need to design your informatica using lookup to find out whether the records exist or not in target table. Since target table has huge volume so it’s costly to create cache it will hit the performance.
* If we create staging tables in the target database, we can simply check records present or not.
* It will avoid full table scan to determine insert or update records.