**Assignment(1)**

**Class:MCA-II(Sem-III)**

**Subject: Business Intelligence**

**Topics:** ETL Process and applications of ETL, Difference Between structured,semi structured and unstructured data?

**Submitted To:**

Manisha Bhatnagar

Prof.of Comp.Sci.Department

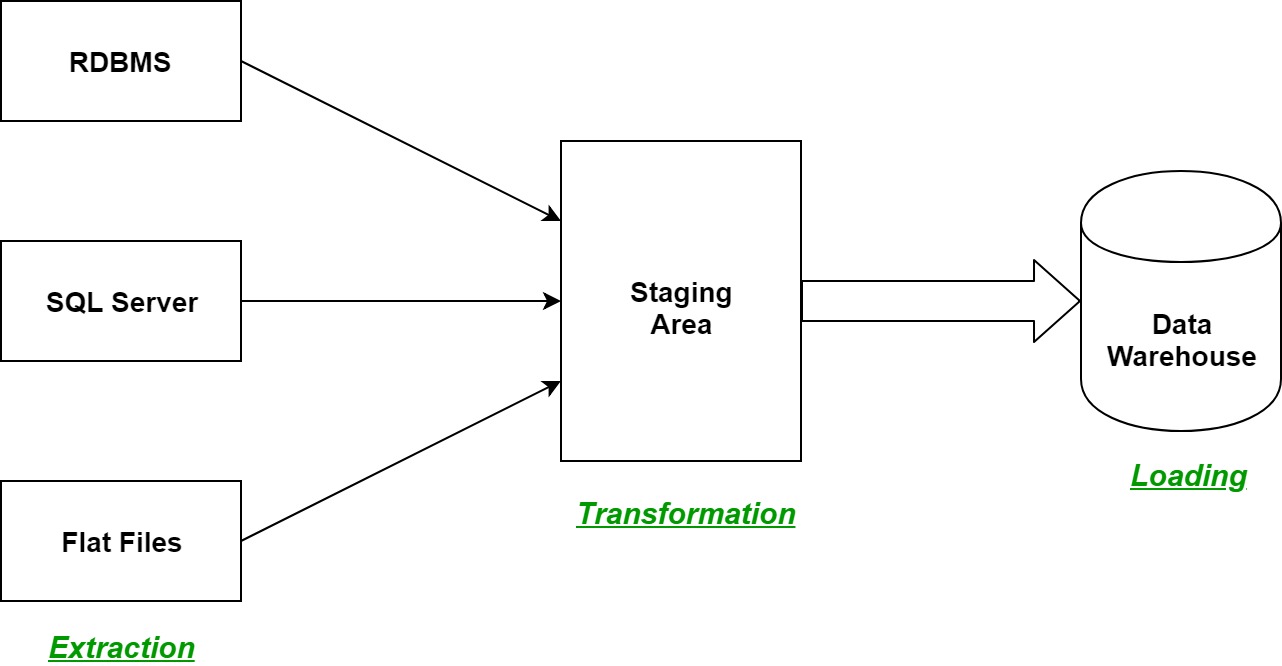
**Ques1:Explain ETL Process and applications of ETL?**

**Ans.** ETL is a process in Data Warehousing and it stands  for Extract**,** Transform and Load**.** It is a process in which an ETL tool extracts the  data from various data source systems, transforms it in the staging area, and then  finally, loads it into the Data Warehouse system.

ETL provides the foundation for data analytics and machine learning workstreams.  Through a series of business rules, ETL cleanses and organizes data in a way  which addresses specific business intelligence needs, like monthly reporting, but it  can also tackle more advanced analytics, which can improve back-end processes  or end user experiences. ETL is often used by an organization to:

• Extract data from legacy systems

• Cleanse the data to improve data quality and establish consistency • Load data into a target database



Let us understand each step of the ETL process in-depth:

1. **Extraction:**

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, No SQL, XML, and flat files 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 and can be corrupted also. 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.

2. **Transformation:**

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

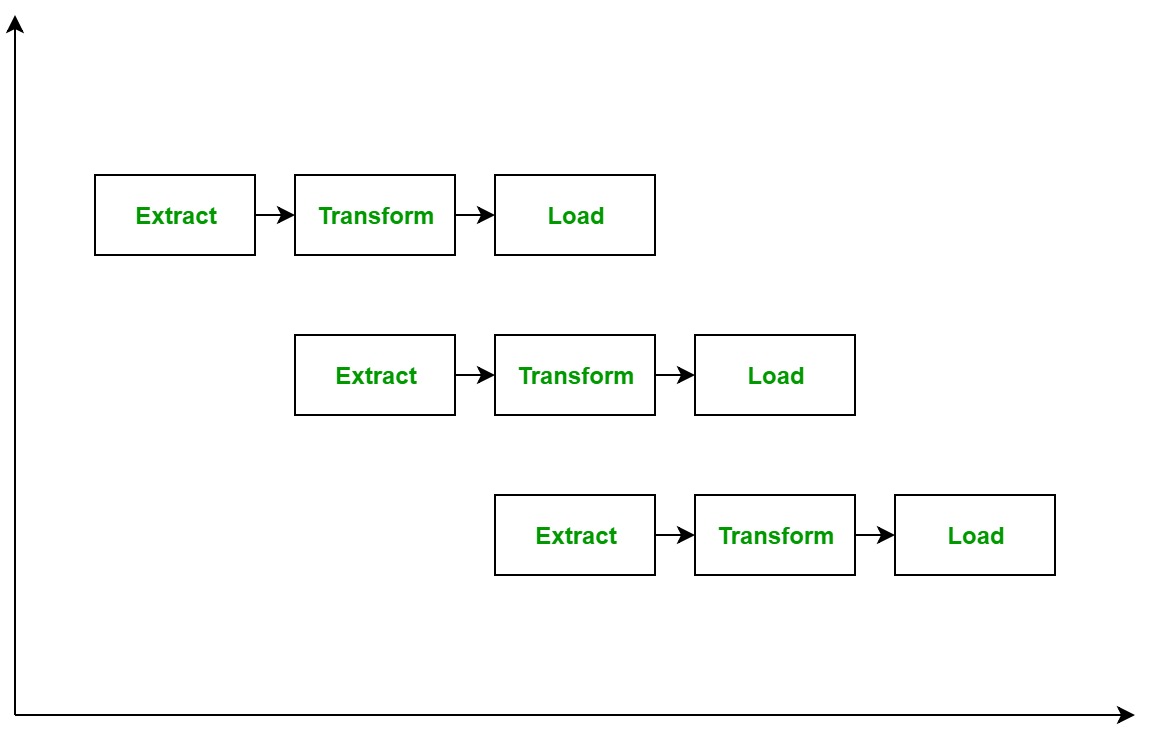
• Joining – joining multiple attributes into one.

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

3. **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.

ETL process can also use the pipelining concept i.e. as soon as some data is  extracted, it can transformed and during that period some new data can be  extracted. And while the transformed data is being loaded into the data  warehouse, the already extracted data can be transformed. The block diagram of  the pipelining of ETL process is shown below:



**Applications of the ETL process are :**

• To move data in and out of data warehouses. Databases are not suitable for big  data analytics therefore, data needs to be moved from databases to data  warehouses which is done via the ETL process.

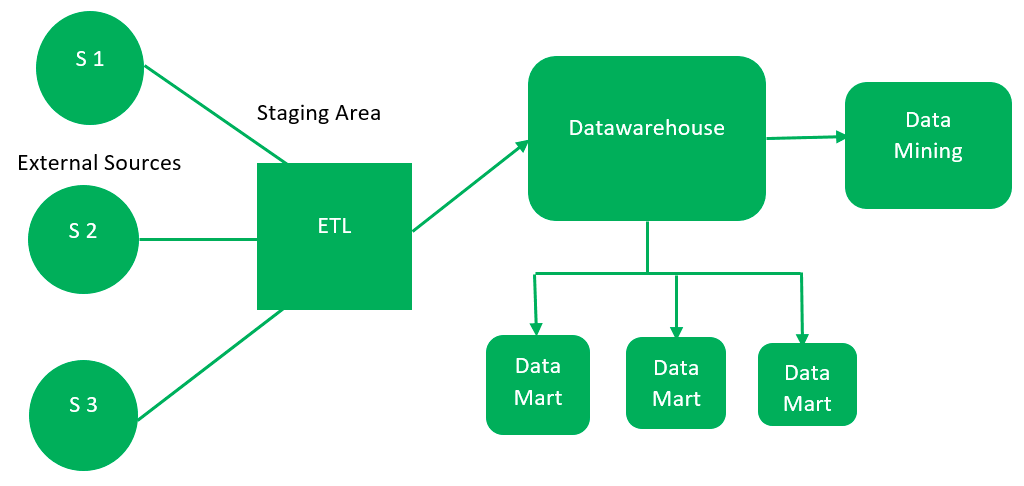
• Data strategies are more complex than they have ever been. ETL facilitates to  transform vast quantities of data into actionable business intelligence.

**There are two approaches in ETL :**

1. **Top Down Approach :**

The data flow in the top-down OLAP environment begins with data extraction  from the operational data sources. This data is loaded into the staging area and  validated and consolidated for ensuring a level of correctness and then moved to  the Operational Data Store (ODS).

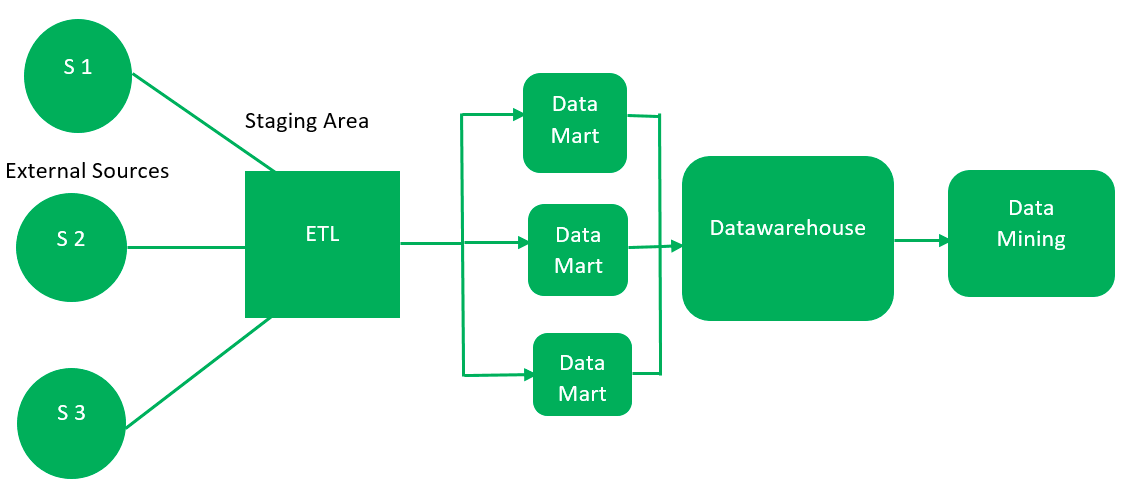
The ODS stage is sometimes skipped if it is another copy of the operational  databases. Data is loaded into the Data warehouse in a parallel to avoid  extracting it from the ODS. Data is routinely extracted from the ODS and  temporarily hosted in the staging area for aggregation, summarization and then  extracted and loaded into the Data warehouse.



1. **Kimball Methodology (Bottom-Up Approach) :**

The bottom‐up approach reverses the positions of the Datawarehouse and the  Data marts. Data marts are directly loaded with the data through the staging  area. The existence of ODS depends on business requirements. The data flow  in the bottom-up approach starts with the extraction of data from operational  databases into the staging area where it is processed and consolidated and then  loaded into the ODS.

The data in the ODS is either appended to or replaced by the fresh data being  loaded. Once the ODS is refreshed, the present data is once again extracted into  the staging area and processed. The data from data mart is pulled to the staging  area aggregated, summarized, and so on and loaded into the Data Warehouse  and made available to the end-user for analysis.



**Advantages of ETL Tools :**

• Easy to use.

• Load data from different targets at same time.

• Performs data transformation as per need.

• Better for complex rules and transformations.

• Inbuilt Error handling functionality.

• Based on GUI and offer visual flow.

• Save Cost and generate higher revenue.

**Disadvantages of ETL Tools :**

• Not suitable for near real-time data access.

• Inclined more towards batch data processing

• Difficult to keep up with changing requirements.

**Ques2:Difference Between structured,semi structured and unstructured data?**

**Ans.**

| **Properties** | **Structured data** | **Semi-structured data** | **Unstructured data** |
| --- | --- | --- | --- |
| Technology | It is based on Relational database table | It is based on XML/RDF(Resource Description Framework). | It is based on character and binary data |
| Transaction management | Matured transaction and various concurrency techniques | Transaction is adapted from DBMS not matured | No transaction management and no concurrency |
| Version management | Versioning over tuples,row,tables | Versioning over tuples or graph is possible | Versioned as a whole |
| Flexibility | It is schema dependent and less flexible | It is more flexible than structured data but less flexible than unstructured data | It is more flexible and there is absence of schema |
| Scalability | It is very difficult to scale DB schema | It’s scaling is simpler than structured data | It is more scalable. |
| Robustness | Very robust | New technology, not very spread | — |
| Query performance | Structured query allow complex joining | Queries over anonymous nodes are possible | Only textual queries are possible |