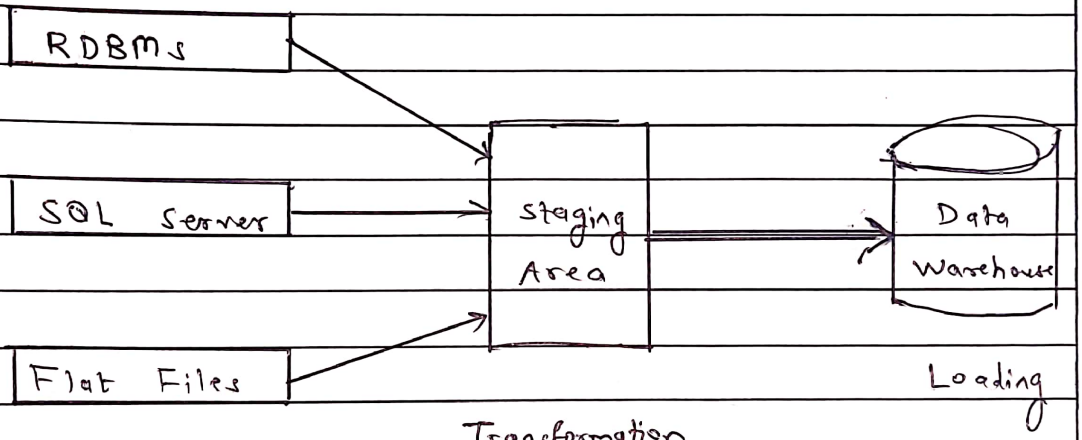
6(A)

ETL Process

- 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.



① Extraction:

- The first step of the ETL process is extraction. In this step data from various source systems is extracted which can be 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

TU3F1819127 **SIGNATURE:** Amey.

staging area first and not directly into the data warehouse because the extracted data is in various formats and can be corrupted also.

(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 tasks.

(A) Filtering
- Loads only certain attributes into the data warehouse.

(B) Cleaning
- Filling up the Null values with some default values, mapping U.S.A., etc.

(C) Joining
- Joining multiple attributes into one

(D) Splitting
- Splitting a single attribute into multiple attributes

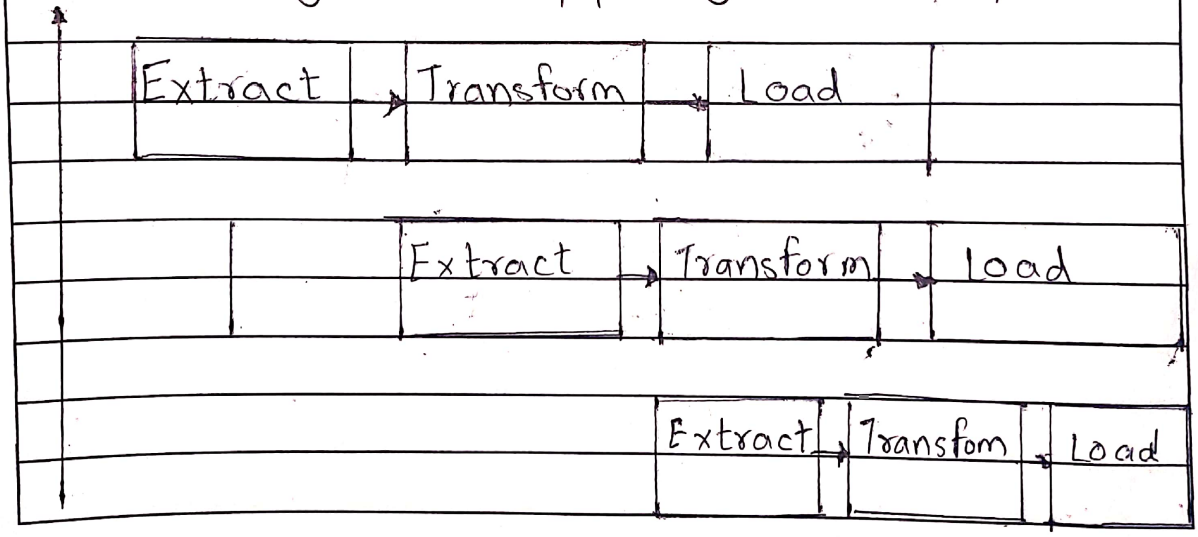
(E) Sorting
- Sorting tuples on the basis of some attribute

③ Loading :

- The third and final step of the ETL proces. is loading in this state the transform data is finally loaded into data warehouse.
- Sometimes 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.

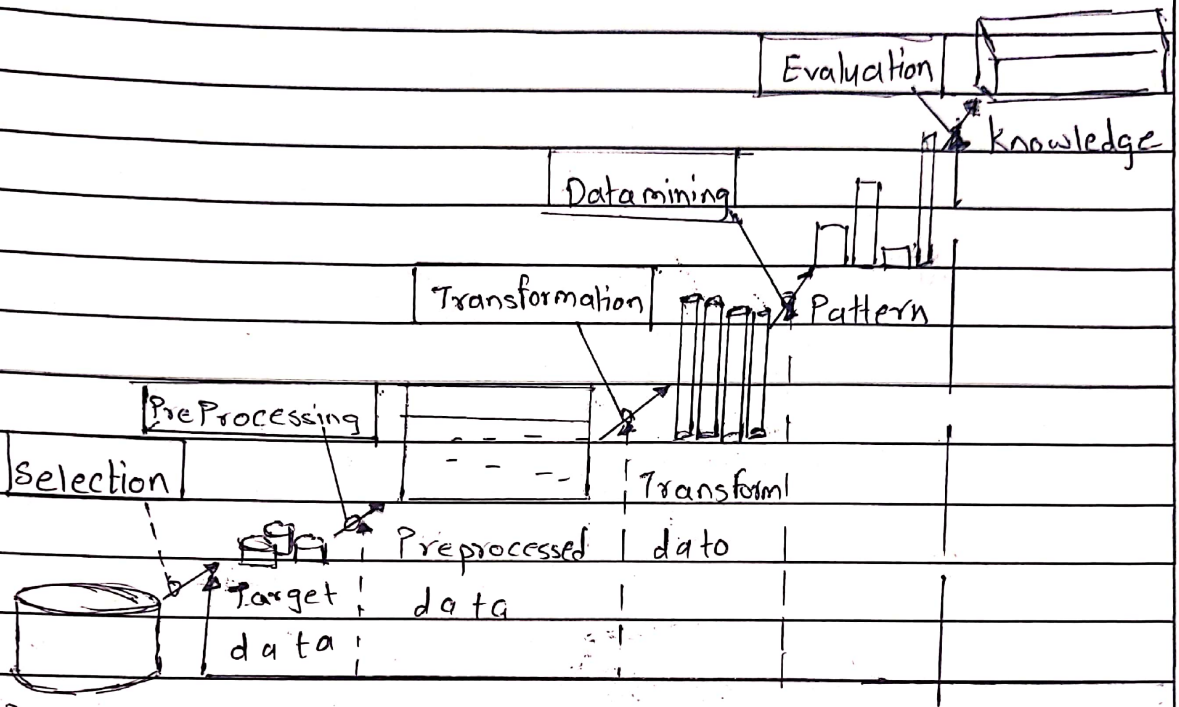
ETL process can also use the pipelining concept.. i.e. as soon as some data is extracted, it can transformed and during that period some new data can be extracted.

Block Diagram of pipelining of ETL process



6 (B)

KDD process



Steps of the KDD Process

The overall process of finding and interpreting patterns from data involves the repeated application of the following steps.

(1) Developing an understanding of
- the application domain
- the relevant prior knowledge
- the goals of the end user

(2) Creating a target data set:
- Selecting a data set, or
- focusing on a subset of variable, or
- data samples, on which discovery is to be performed.

(3) Data cleaning and preprocessing
- Removal of noise or outliers.
- Collecting necessary information to model or account for noise.
- Strategies for handling missing data fields.
- Accounting for time sequence information and known changes.

(4) Data Reduction and Projection.
- Finding useful features to represent the data depending on the goal of the task.
- Using dimentionality reduction or transformation methods to reduce the effective number of variable under consideration or to find invariant representation for data.

(5) Choosing the data mining task
- Deciding whether the goal of the KDD process is classification, regression, clustering, etc.

(6) Choosing the data mining algorithms
- Selecting methods to be used for searching for patterns of the data
- Deciding which models and parameters may be appropriate
- Matching a particular data mining method with the overall criteria of the KDD process

(7) Data mining
- Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering and so forth

(8) Interpreting mined patterns

(9) Consolidating discovered knowledge