Shantanu S. Mangale
BE. B. 41242 DMW
LP-II

Assignment No.1

Title:- For an organization of your choice, choose a set of business processes. Design star/snowflake schemas for analyzing these processes. Create a Fact constellation schema by combining them. Extract data from different data sources, apply suitable transformations & load into destination tables using an ETL tool.

Problem Definition:-
Design a basic ETL model using Rapid Miner Application.

Pre-requisite:-
i) Basic concepts of ETL.
ii) Knowledge about Rapid Miner tool.

S/W & H/W Requirements:-
Rapid Miner, PIV, 2GB RAM, 500 GB HDD.

Learning Objectives:-
Understand the implementation of the various ETL model using RapidMiner Tool.

Outcomes:- After completion of this assignment students can develop & analyze the ETL model & will understand the working.

Theory :-

- What does ETL mean?

ETL stands for Extract - Transform~~atio~~ & Load.
An ETL tool extracts the data from different RDBMS source systems, transforms the data like applying calculations, concatenate, etc. & then load the data to Data Warehouse system.

- Extraction:

i) A staging area is required during ETL load. There are various reasons why staging area is required.

ii) The source systems are only available for specific period of time to extract data. This period of time is less than the total data-load time.

iii) Staging area is required when you want to get the data from multiple data sources together or if you want to join two or more systems together.

iv) Data extractions' time slot for different systems vary as per the time zone & operational hours.

- Transform :-

In data transformation, you apply a set of functions on extracted data to load it into the target system. Data, which does not require any transformation is known as direct move or pass through data.

You can apply different transformation on extracted data from the source system.

- Load

During Load phase, data is loaded into the end-target system & it can be a flat file or a Data Warehouse system.

- Tool for ETL : Rapid Miner :-

Rapid Miner is a world-leading open-source system for data mining. It is available as a stand-alone application for data analysis & as a data mining engine for the integration into own products. Rapid Miner is now Rapid Miner Studio & Rapid Analytics is now called Rapid Miner Server.

- Data Warehouse Schemas:-

i) Star schema.

ii) Snowflake schema.

iii) Fact Constellation.

- Star Schema:-

For eg:- as you can see in the above-given image that Fact table is at the centre which contains keys to every dimension table.

→ Characteristics of Star Schema:

i) Every dimension in a star schema is represented with the only one-dimension table.

ii) The dimension table should contain the set of attributes.

iii) The dimension table is joined to the Fact table using a foreign key.

iv) The dimension table are not joined to each other.

v) Fact table would contain key & measure.

- Snowflake Schema:

A snowflake schema is an extension of a star schema & it adds additional dimensions. It is called snowflake because its diagram resembles a snowflake.

→ Characteristics of Snowflake Schema:

i) The main benifit of the snowflake schema it uses smaller disk space.

ii) Easier to implement a dimension is added to the Schema.

iii) Due to multiple tables query performanre is reduced.

Conclusions:-

With the help such Tools we can Perform ETL operations on Sample Data sets & can perform analysis on sample data sets.