Date
Page

LP2 - Assignment A1

Title:- Data mining

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

Objective:-
1) Study diff. schemas for data modelling.
2) Understand basic principles of data mining.

Outcome:- Student will be able to extract useful info from different data sources.

S/W & H/W :-
1) 64 bit OS
2) Pentaho
3) 4 GB RAM

Theory:-

Star schema:- The most common modelling paradigm is the star schema, in which the data warehouse contains (1) Large central data table containing the bulk of the data with no redundancy & (2) a set of smaller attendant table one for each dimension.

The schema graph resembles a starboat with dimensions of table displayed in a radial pattern around the centre fact table.

Snowflake schema:- It is the variant of star schema model where so dimensional table are normalized, thereby for the further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake.

Fact constellation:- Sophisticated applications may require multiple fact tables to share dimension tables. This kind of schema can be viewed as a collection of stars & hence is called a galaxy or a fact constellation.

ETL:- It is an abbreviation for Extract, Transform & Load. In this process, an ETL tool extract the data from diff. RDBMS source systems then transforms the data like applying calculations, concatenations, etc. and then load the data into various Data warehouse System.

Pentaho:- With an intuitive, graphical, drag & drop environment & a proven scalable, standards based architecture, Data integration is increasingly the choice for org. over traditional ETL tools.

Algorithm:-

1) Start Pentaho.
2) Create a new transformation.
3) Retrieve data from a flat file
4) Filter records
5) Load data from lookup file.
6) Complete transformation

Conclusion:- Thus, we have learnt to extract data from different data sources, apply suitable transformations & load into destination tables using an ETL tool.