Assignment : DMW 1.

Aim: for an organization of your choice chouse set of attributes business process. Design star/ snow flake schemas for analyzing these processes. Create a fact constellation scheme by combining them. Extract data from different data sources apply suitable transformations & load into destination tables using an ETL Tool for Ex: Business organization: sales, order Marketing processes

Objectives :-
understand basics of star / snow flake / fact constellation schema & learn the rapid miner tool for performing various operation on in built or external datasets.

Outcomes:
i) Students will be able to demonstrate installation of rapidminer tool.
ii) Students will be able to demonstrate diff operator and datasets in rapidminer
iii) Students will be able to demonstrate different operations on available data in rapid miner.

H/w Requirement :- Any CPU with pentium processor or similar.

S/w Requirement: ETL tool software, OS.

Teacher Signature

Theory:-

What does ETL mean?

ETL stands for Extract, Transform and load. An ETL tool extracts data from different RDBMS source systems transforms the data like applying calculations, concatenation etc & then load the data to data warehouse system. The data is loaded in Data Warehouse System in the form of dimension and fact tables.

Extraction:-

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

The source systems are only available for specific time to extract data This period of time is less ~~tan~~ than total data load.

Staging area is required when you want to get data from multiple ~~re~~ data sources together or if you want to join two or more systems together

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

Teacher Signature _____

ETL allows to perform various complex transformations and requires extra area to store data.

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 tranformations on extract data from the source system for ex. you can perform customized calculations. If you want sum of sales revenue and this is not in data base you can apply sum formula during transformation

Load.

During load phase data is loaded into the end target system and it can be a flat file or a data warehouse system.

Data Warehousing System schemas.
1. Star scheams.
2. snowflake schemas
3. fact constellation.

Teacher Signature

STAR SCHEMA	SNOWFLAKE SCHEMA
2) Hierarchies for the dimensions are stored in the dimentional table	1) Hierarchies are divided into separate table.
2) It contains a fact table surrounded by dimention table.	2) One fact table surrounded by dimension table. which are in turn surrounded by dimention table
3) Simple DB Design	3) Complex DB Design
4) In a star schema only single join creates the relⁿ b/w fact table & any dimentional tables.	4) A snowflake schema requires many joins to fetch the data
5) High level of Data redundancy	5) very low level Data Redundancy
6.) Denormalized data structured query also run faster	6) Normalized data structured
7) Single dimentional data table contains aggregated data	7) Data split into diff Dimention tables.

Conclusion:- Hence we are able to study rapidminer tools can perform ETL operations on sample datasets and can perform analysis on simple datasets.

Teacher Signature _____