Data Warehouse & Data Mining

Assignment - 2

Ques 1) What is data warehouse? Explain the characteristics of data warehouse.

Ans- Data warehouse is a large store of data accumulated from a wide range of sources within a company and used to guide management decisions.
For effectively performing analytics, an organization keeps a central data warehouse to closely study its business by organizing, understanding and using its historic data for taking strategic decisions and analyzing trends.

Eg applications of data warehousing

i) Social media websites - They gather data for analyzing user's interests
ii) Banking - They analyze the spending patterns of account/ cardholders may be to provide them with special offers.
iii) Government - They store and analyze tax payments which are used to detect tax thefts.

Characteristics of data warehouse

i) Subject oriented -
Data warehouse is subject oriented because it delivers information about a theme or a particular subject like sales, distributions, marketing etc instead of organization's current operations.
This subject focuses on demonstrating and analysis of data to make various decisions.

ii) Integrated -
Integration means founding a shared entity to scale all similar data from the different databases.
A data warehouse is built by integrating data from various sources of data such that a mainframe and a relational database. It benefits effective analysis of data.

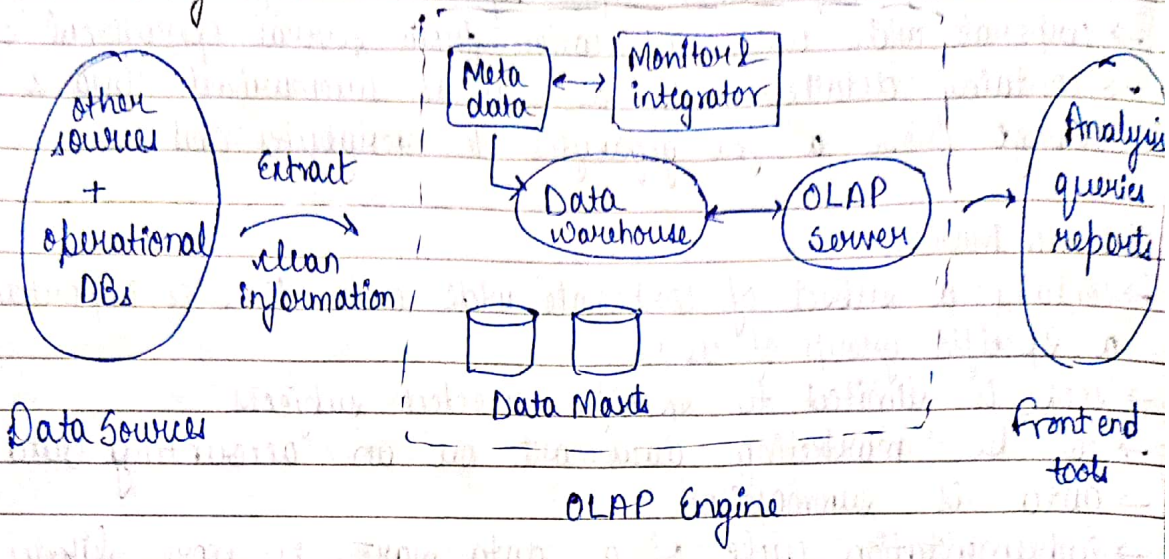
iii) Time variant -
In this, data is maintained via different intervals of time such as weekly, monthly or annually. The time limits for data warehouse is wide ranged than that of operational systems.
It is also time-variance, i.e, once data is stored in the warehouse then it cannot be modified, alter or updated.

iv) Non - volatile -
The data in the warehouse is permanent. It means that the data is not erased or deleted when new data is inserted. It evaluates the analysis within the technologies of warehouse.
Only two operations are done in the data warehouse.
→ Data Loading
→ Data Access

Benefits of Data Warehouse
i) easy structure for easy navigation and understanding
ii) complex queries become simpler
iii) managing data from different and scattered sources is easier
iv) efficient to manage demand for lots of information
v) ability to analyze large amounts of historical data.

Ques 2) Describe the three-tier architecture of data warehouse with diagram.

Ans-



i) Bottom Tier (Data sources & data storage)
→ usually consists of data sources
→ warehouse database server
→ In bottom tier, using API's (gateways), data is extracted from operational and external sources

ii) Middle Tier
→ OLAP server
→ It enables users to easily and selectively extract and query data in order to analyze it from different points of view.

iii) Top Tier
→ front-end client layer
→ includes query and reporting tools, analysis tools and/or data mining tools for trend analysis, prediction etc.

* From the architecture point of view, there are three warehouse models.

i) Enterprise Warehouse
→ collects all information topics spread throughout the organization
→ corporate-wide data integration from several operational systems
→ contains detailed data as well as summarized data & can range from a few gigabytes to terabytes and more.

ii) Data Mart
→ contains a subset of corporate-wide data that is important to a specific group of users.
→ scope is limited to specific selected subjects
→ can be marketing datamart or an accounting data mart
→ data is summarized
→ implementation cycle of a data mart is more likely to be measured in weeks

iii) Virtual warehouse
→ group of views on an operational database
→ Creating it is easy but requires additional capacity on operational database servers.

Que3) Discuss the multi dimensional data model. Also determine schemas for multi dimensional databases.

Ans- Multidimensional data models are designed expressly to support data analysis.
→ Data warehouses and OLAP tools are based on a multidimensional data model.
→ This model views data in the form of "data cube"
→ Cube is a 3D structure which makes the data retrieval efficient.

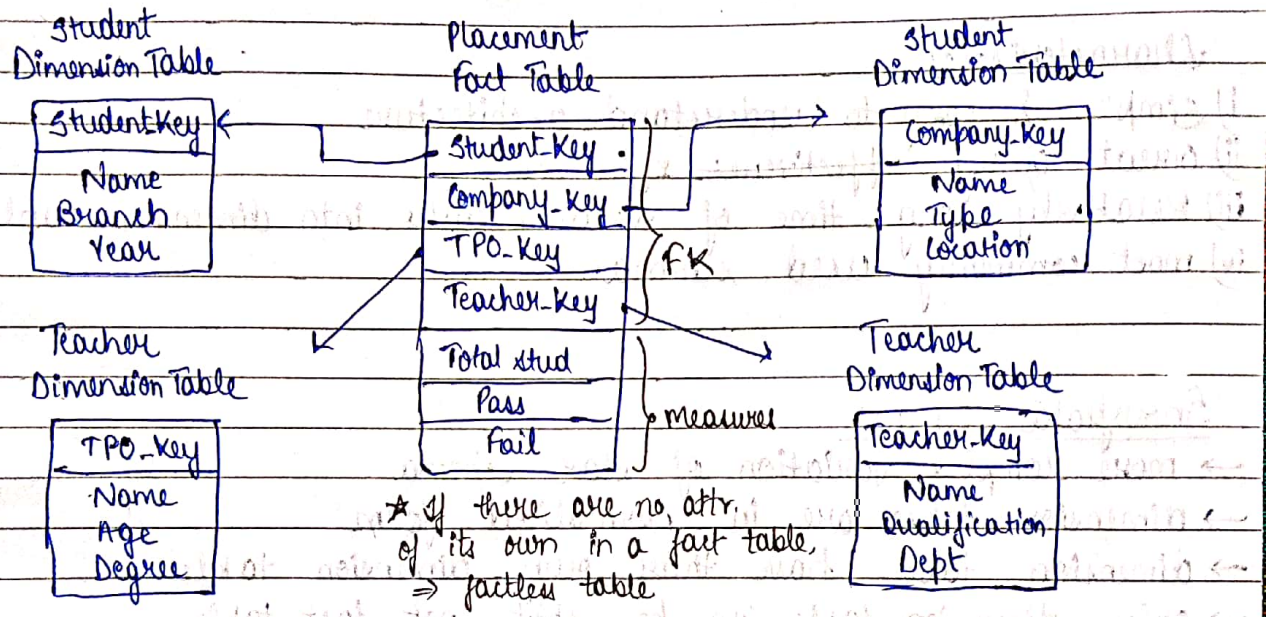
Types of multidimensional model

- Star Schema
- Snowflake Schema
- Fact Constellation Schema

Schema

It is a logical description of entire data database.

Fact Table



A fact table also called as master table typically has two types of columns: foreign keys to dimension tables and measures those that contain numeric facts

A fact table can contain fact's data on detail or aggregated level.

Dimension Table

A dimension is a structure usually composed of one or more hierarchies that categorizes data. If a dimension hasn't got a hierarchy & levels, its called flat dimension or list.

→ The primary keys of each of the dimension tables are part

of the composite primary key of the fact table. Dimensionless attributes help to describe the dimensionless value.

Star Schema

→ shape of a star with points radiating from the center
→ center = fact table and points = dimension tables
→ The fact table usually is in third normal form and dimension tables are denormalized

Characteristics

i) Simple & easy to understand architecture
ii) Great query ~~effectiveness~~ effectives
iii) Relatively long time of loading data into dimension tables
iv) most commonly used schema

Snowflake Schema

→ more complex variation of star schema
→ dimension tables are in normalized form
→ dimension table have their own dimension table
→ only dimension table can be split, not fact table
→ Redundancy & space are reduced due to normalized table

Fact Constellation Schema

→ combination of star and snowflake schema
→ also called Galaxy schema
→ contains multiple fact tables that share many dimension tables
→ complex design & large dimension tables.