Data warehouse Architecture

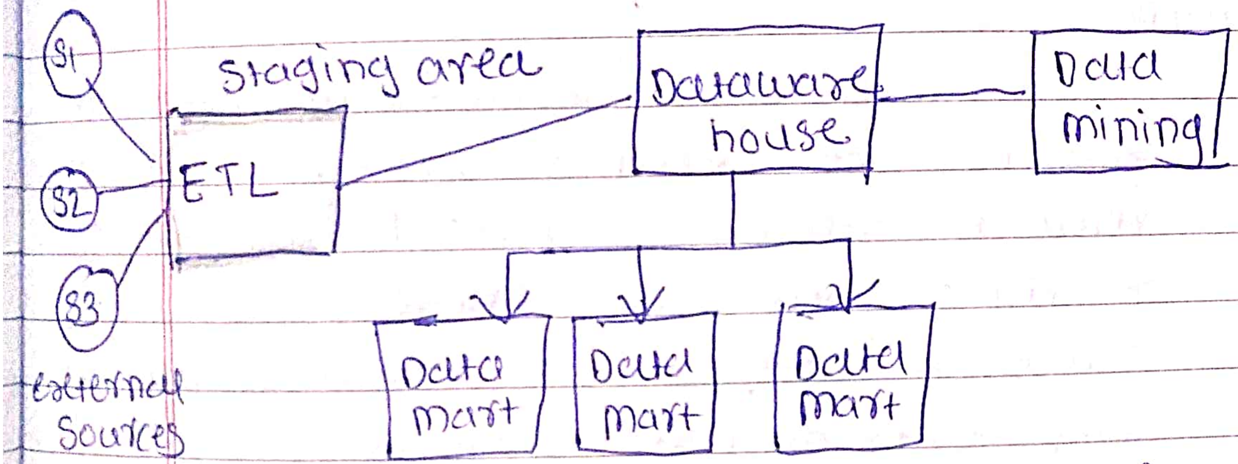
Intro:-

A data warehouse is a heterogeneous collection of different data sources. organized under a unified schema. There are two approaches

i) Top-down approach.

ii) Bottom-up approach.

Top-down



The essential components are as follows:-

i) External Sources - It is a source from where data is collected irrespective of the type of data.

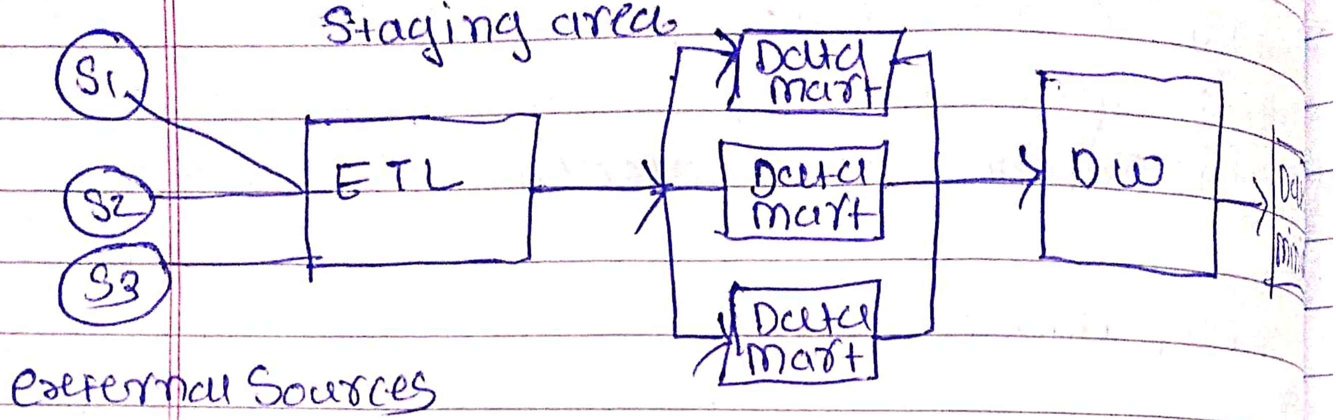
ii) stage area :- Since the data, extracted from the external sources does not ~~follo~~ follow a particular format, so there is a need to validate this data

iii) Data warehouse - After cleansing of data it is stored in the data warehouse as central repository.

iv) Data Marts - It is also a part of storage component.

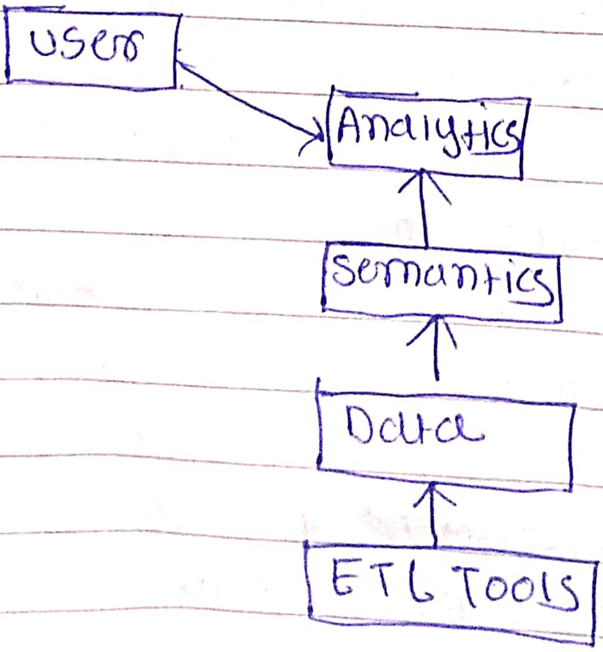
Data mining - It is a practice of analysis the big data present in datawarehouse

Bottom-up approach -



Data is extracted from external sources' Then data go through the staging area & loaded into Data Marts instead of data warehouse. These data marts are then integrated into datawarehouse

components of Datawarehouse Architecture



A typical data warehouse include the three seperate layers above

Data layer - Data is extracted from your sources & then transformed & loaded into the bottom tier using ETL Tools. It consist of database server, datamarts & data lakes. Metadeta is created in this ~~ty~~ tier.

semantics layer. In the middle tier, OLAP & OLTP servers restructure the data for fast, complex queries & analytics.

Analytics layer:- The top tier is the front end client layer. It holds the data warehouse access tools that let users interact with data, create dashboards & reports. This

Federated data warehouse:-

It is a practical approach to achieving the "single version of the truth" across the organization. It is used to integrate key business, measures & dimensions.

Architecture:-

A big organization has various regions that provide business to customers globally. Diff. regional data warehouses were built for each region to meet the specific business needs. Diff. bet^{n} regional & global data warehouse system is the nature of data resided at each system level. There are two data flow between regional & global data wareho upward Federation - only fact data ar moved from regional data warehouse to global data warehouse.

Downward federation - The reference flow from global to the regional level.

Dimensional modeling :- It represents data with cube operation, making more suitable logical data representation with OLAP data management. In dimensional modeling the transaction record is divided into either "facts" whi are frequently numerical transaction data or dimensions which are the reference information that gives context to the facts. The purpose of dimensional modeling are to produce database architecture that is easy for end-clie to understand & write queries. To maximize efficiency queries.

Difference between ER modeling & Dimensional modeling

ER data modeling	Dimensional Data modeling
suitable for OLTP & Application	suggested for Data warehouse application.

Consist of Entities & Relationships	Consist of Facts & Dimensions.
High CRUD Activity	High select activity.
Normalization is suggested	De-normalization suggested.

Data warehouse Schemas.

Schema is a logical description of the entire database. It includes the name & description of records.

Star Schema :- It is the elementary form of a dimensional model, in which data are organized into facts & dimension.

A fact is an event that is counted.

A dimension includes reference data about fact. It is a relational schema.

Dimensional Table :- A dimension is an architecture usually composed of one or more hierarchies that categorize data. Dimensional attributes help to define the dimensional value.

Snowflake Schema :-

Some dimension tables in the Snowflake schema are normalized. The normalization splits up the data into additional tables.

Fact Constellation Schema :-

It has mutiple fact tables. It is also known as galaxy schema.

Factless Fact table:- These tables are only used to establish relationships bet^{n} ele. of different dimension. They have abbreviated key.

Granularity:- The first step in designing a fact table is to determmine the granularity of fact table. It is a lowest level of information that will be stored in fact table.

It include which dimension will be included determine the hierarchy of each dimension the information will be kept.

metadata:-

It is a data about data. e.g index of a book serves as a metadata for the contents in the book. It is a roadmap to datawarehouse. It defines the warehouse object. It act as directory.

categories of metadata:-

i) Business metadata - It has the data owenership information, business definition & changing policies.

ii) Techinal Metadata - It includes database system names, table & column names & sizes, data types & allowed values.

iii) operational metadata - It includes currency of data & data lineage.

metadata management :- It helps in driving the accuracy of reports.
It has some challenges.
metadata in a big organization is scattered across the organization.
It could be present in text files
There are no easy & accepted methods of passing metadata. metadata is controlled by metadata repository.

metadata management tools :-
There are many tools used to manage the metadata & make the information readily available to the user. metadata management tools help to know the data well & to manage them according to the user's need.
If the data is not managed well, it will be difficult to trace the data.

Types of metadata management tool :-
i) Collibra tool
ii) Alation tool
iii) Infosphere information (IBM tool)
iv) Informatica