

Data Warehousing



What is Data Warehouse

- Data Warehouse is database – organized collection of data.
- *A data warehouse is a central repository of information that can be analyzed to make more informed decisions.*
- Is process of constructing , using a data warehouse.
- Data Warehouse provides
 - a) Generalized data view
 - b) Consolidated data viewwith multi dimensions with tools for OLAP

Purpose of OLAP tools

For

Interactive & effective Analysis for data mining

Data mining functions -

- 1) Association
- 2) Clustering
- 3) Classification
- 4) Prediction

Separation from Operational Database

Data Warehouse is separate from organizations operational or OLTP database

As frequency of updates is less in OLAP database – data warehouse compared to OLTP

Consists of Consolidated Historical data for Analyzing – BI(Business Intelligence)

Data Warehouse - BI

Helps businesses -

a) Organize

b) Understand

c) Use

data to take strategic decisions.

OLAP & OLTP Separate ?

OP – constructed for defined tasks, applications, retrieve few records with indexes etc.

DW -Queries are complex and retrieve general form of data

OP – Concurrency for multiple transactions support is needed with recovery mechanisms – Strict ACID compliance

DW – need to query for analysis so read only access for stored data is sufficient

OP – Query reads and also modifies the data

Features of Data Warehouse

Subject Oriented – information related with specifics like sales

Integrated – Created by various heterogeneous sources –
csv,tsv,ssv,flat files.

Time variant – Particular time period, historical point of view.

Non-Volatile – It's growing database, previous data is not erased.

Data warehouse transactions ?

Data warehouses do not require

- 1) Transactions processing ,
- 2) Recovery,
- 3) Concurrency controls

as frequency of updates is very low.

Applications of Data Warehouse

Domains

- a) Banking
- b) Financial
- c) Consumer patterns, goods
- d) Retail sectors

Types of DW

Information Processing – Querying for Statistical Analysis, reporting using tables, charts, graphs etc.

Analytical Processing – Analyzed by slicing-dicing, drill down, drill up, pivoting.

Data Mining – Knowledge discovery by revealing hidden patterns, trends, associations, constructing models, classification, prediction, mining, visualization etc.

Differences DW & OP

Data Warehouse <i>OLAP</i>	Operational Database <i>OLTP</i>
It involves historical processing of information.	It involves day-to-day processing.
OLAP systems are used by knowledge workers such as executives, managers, and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
It is used to analyze the business.	It is used to run the business.
It focuses on Information out.	It focuses on Data in.
It is based on Star Schema, Snowflake Schema, and Fact Constellation Schema.	It is based on Entity Relationship Model.

It focuses on Information out.	It is application oriented.
It contains historical data.	It contains current data.
It provides summarized and consolidated data.	It provides primitive and highly detailed data.
It provides summarized and multidimensional view of data.	It provides detailed and flat relational view of data.
The number of users is in hundreds.	The number of users is in thousands.
The number of records accessed is in millions.	The number of records accessed is in tens.
The database size is from 100GB to 100 TB.	The database size is from 100 MB to 100 GB.
These are highly flexible.	It provides high performance.

Data Warehousing ?

Datawarehousing is the process of constructing and using a data warehouse.

A datawarehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or adhoc queries, and decision making.

Data warehousing involves data cleaning,data integration,and data consolidations.

Uses of DW

Decision support technologies need data in data warehouse

a) Products and Production Strategies – based on the sales quarter or yearly patterns

b) Customer churn Analysis

c) Operations Corrections Analysis – CRM issues corrections by analyzing business operations

Integration Approach

Query driven – building wrappers and integrators as layer on top of Multiple heterogeneous databases.

Update driven - Information from multiple heterogeneous sources are integrated in advance and stored in databases, support direct querying and analysis.

Tools & Utilities

Data Extraction

- Gathering data from multiple heterogeneous sources.

Data Cleaning

- Finding and correcting errors

Data Transformation

- Legacy to warehouse formats

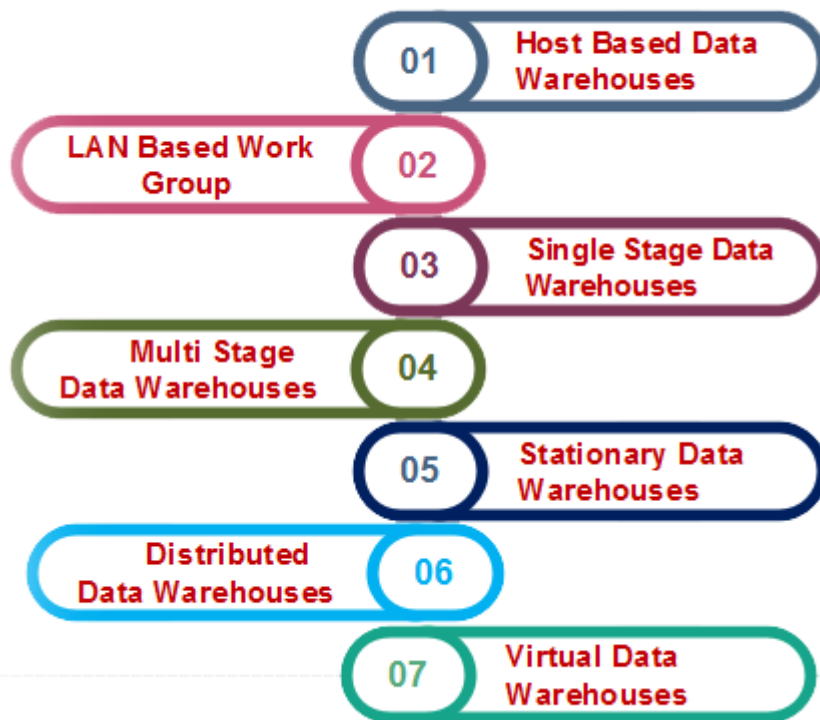
Data Loading

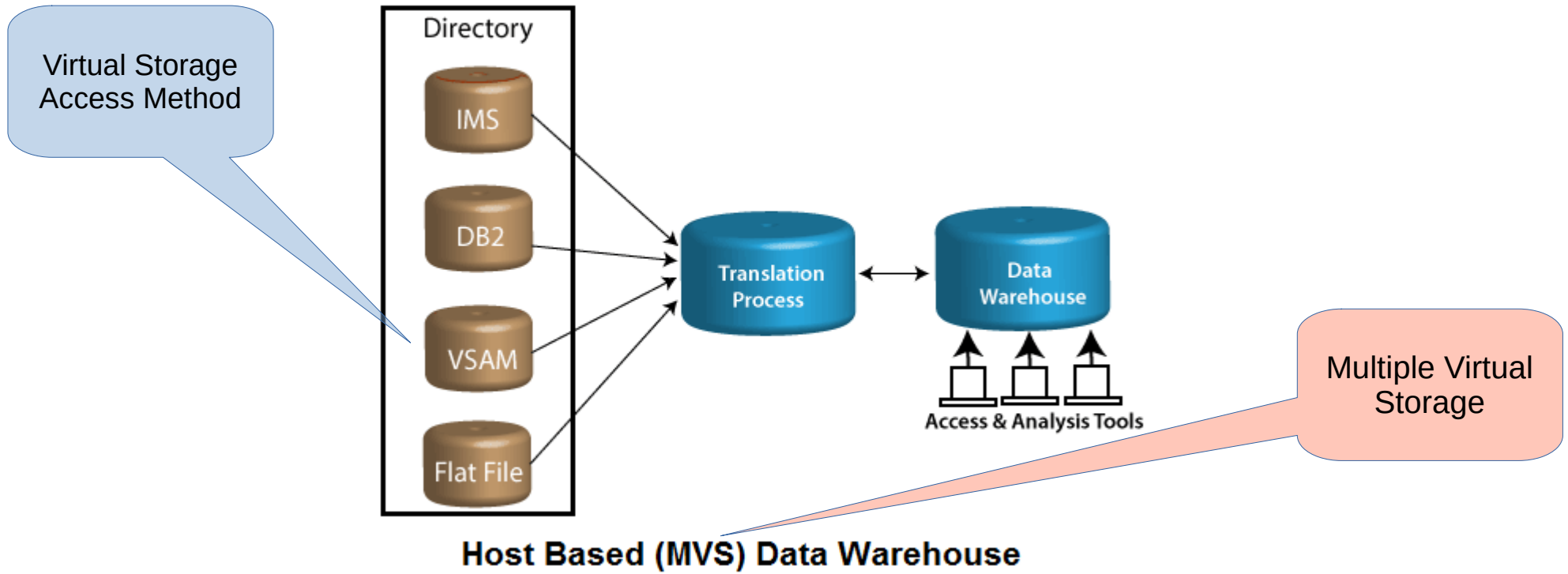
- sorting, summarizing, Consolidating, checking integrity ,build indexes,partitioning etc.

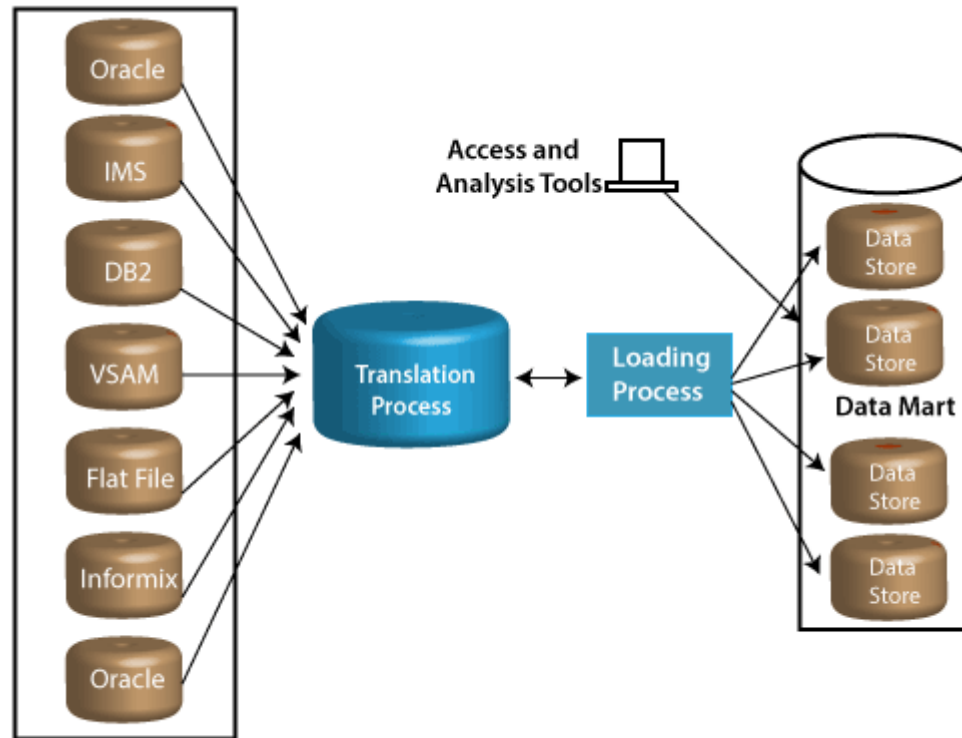
Updating

- Updating from data sources.

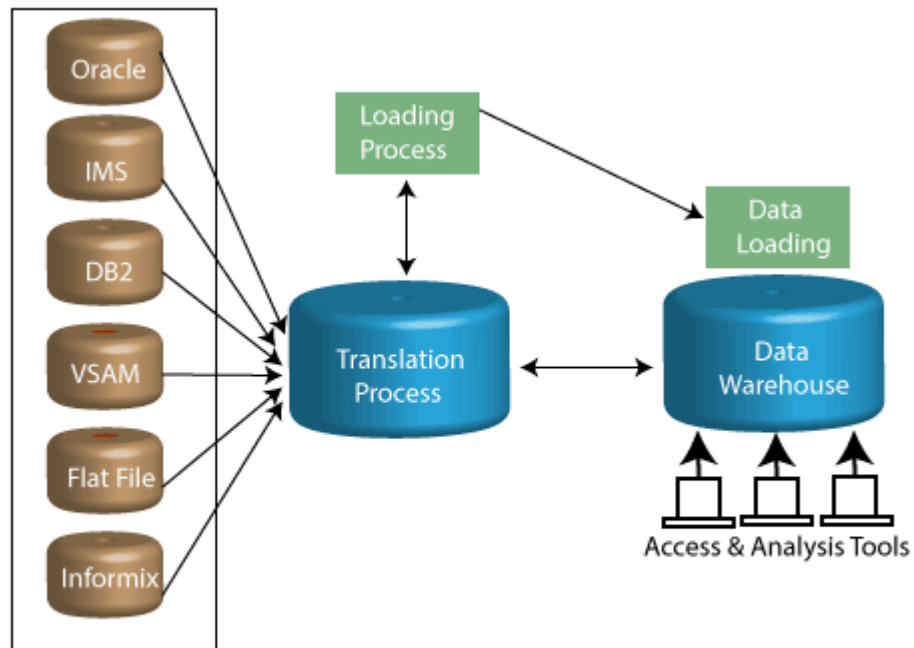
Types of Data Warehouses



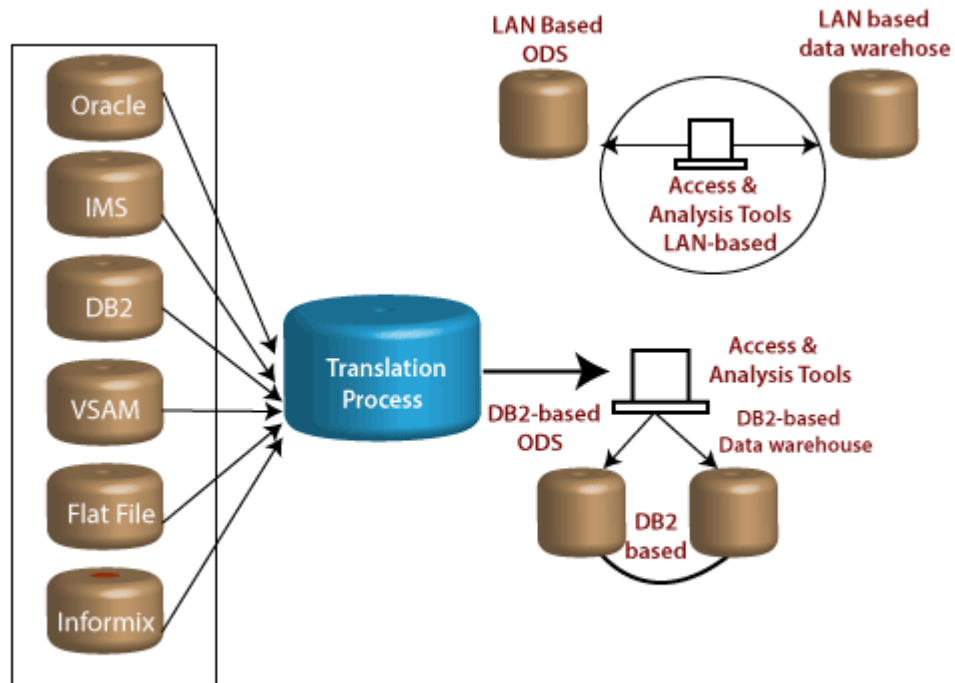




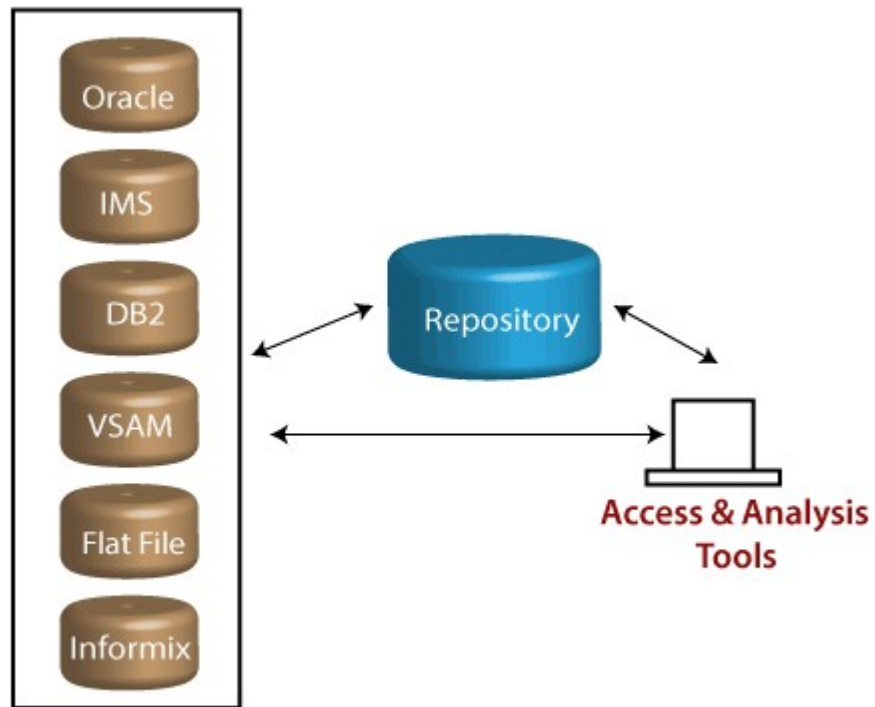
LAN Based Work Group Warehouse



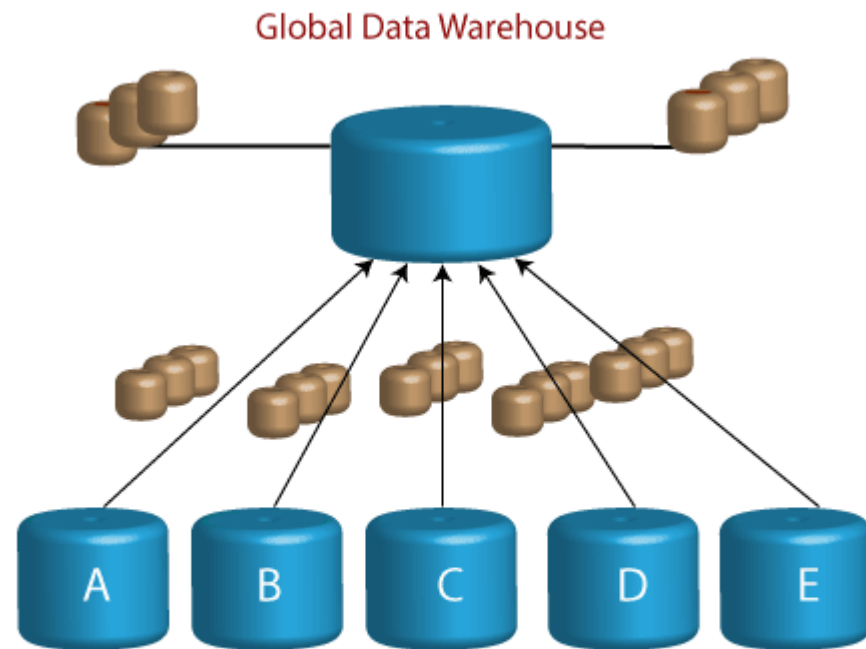
LAN Based Single Stage Warehouse



Multistage Data Warehouse



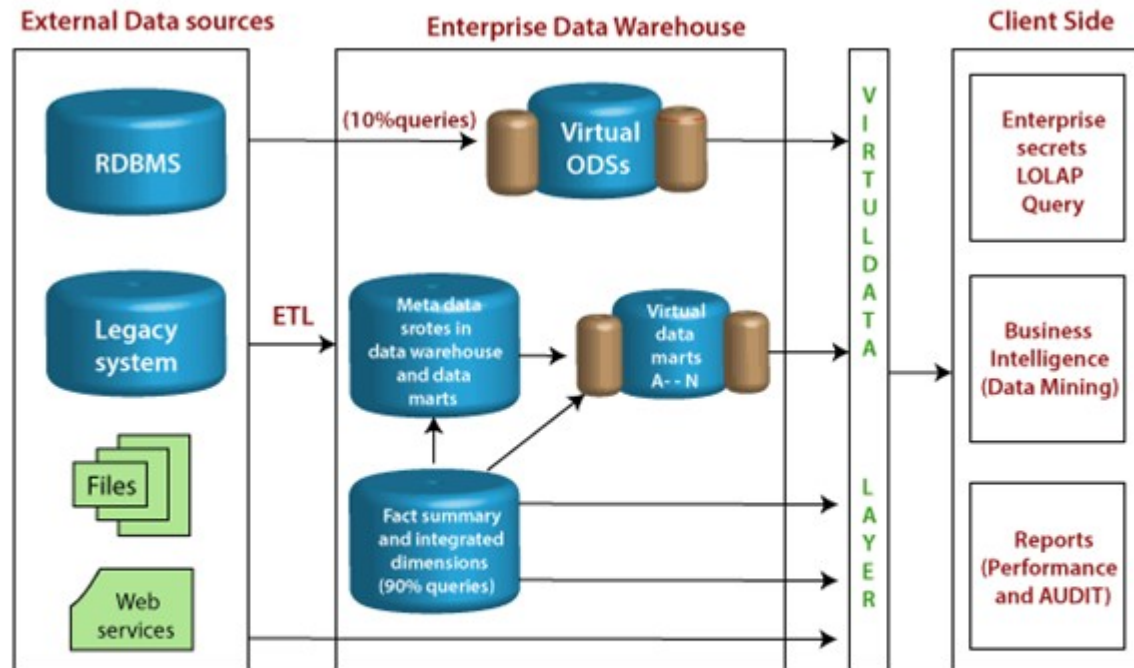
Stationary Data Warehouse



Global Data Warehouse

Local Data Warehouse

Distributed Data Warehouse



10% of user queries are fired on fact summary & 90% of user queries are fired on ODSs

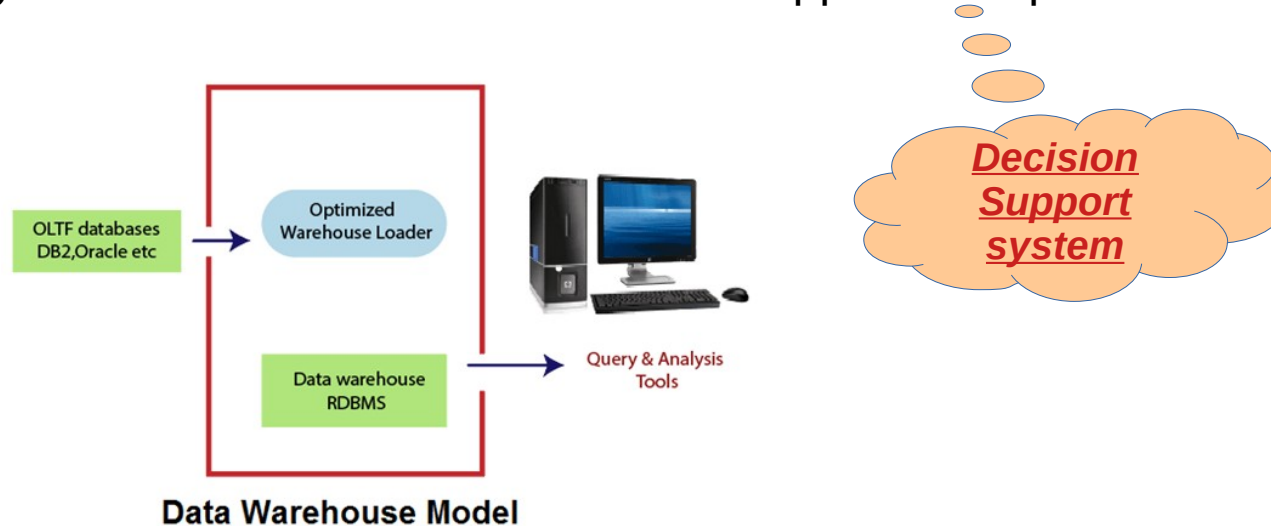
Virtual Data Warehouse

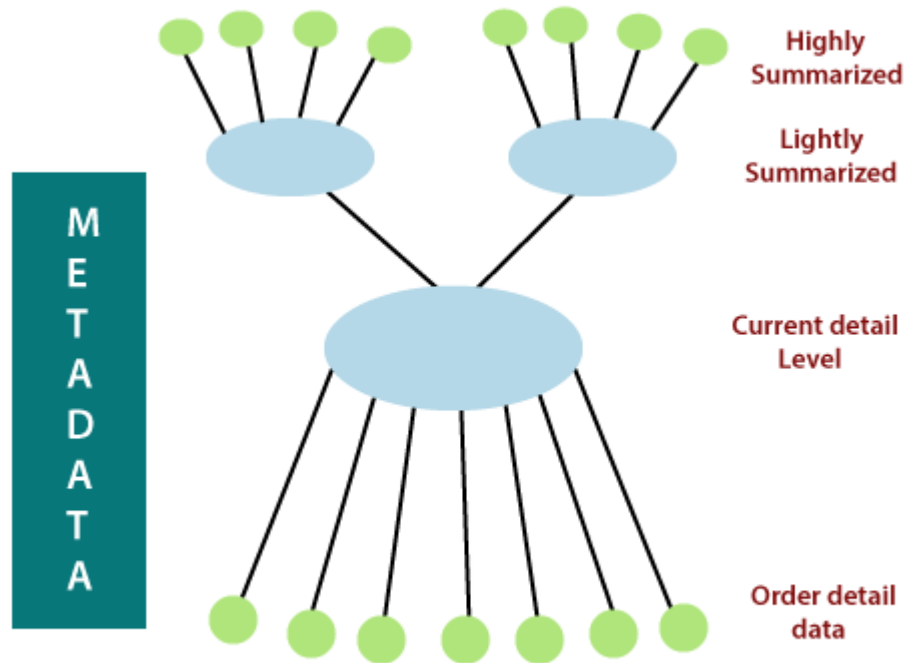
DW Modeling

Data warehouse modeling is the process of designing the schema's of the detailed and summarized information of the data warehouse.

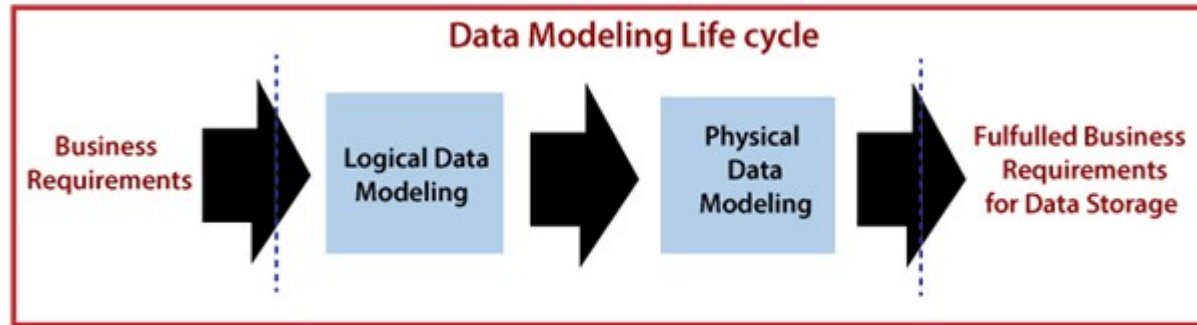
Purpose is data warehouse modeling is to develop a schema describing the reality.

Primary function of data warehouses is to support **DSS** processes.

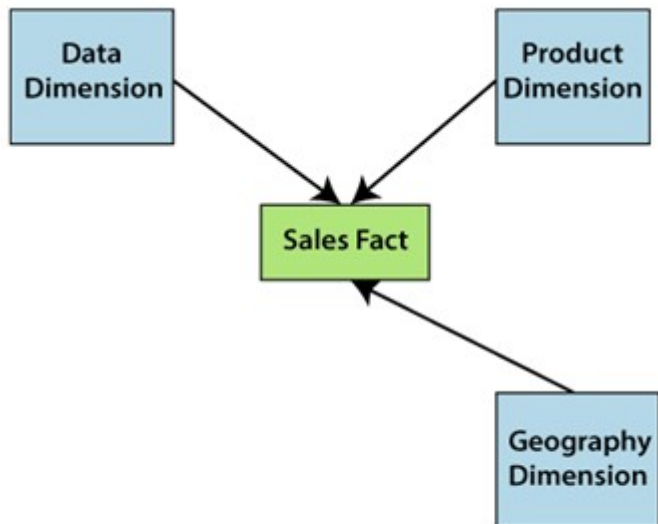




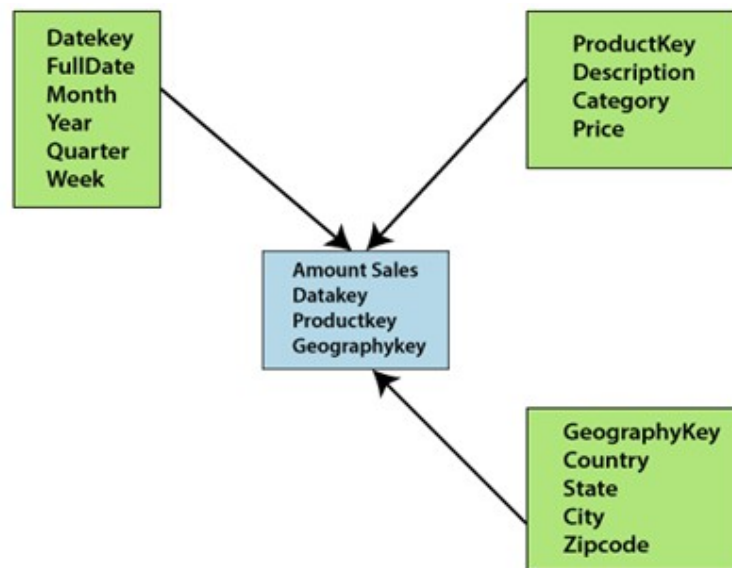
The Structure of data inside the data warehouse



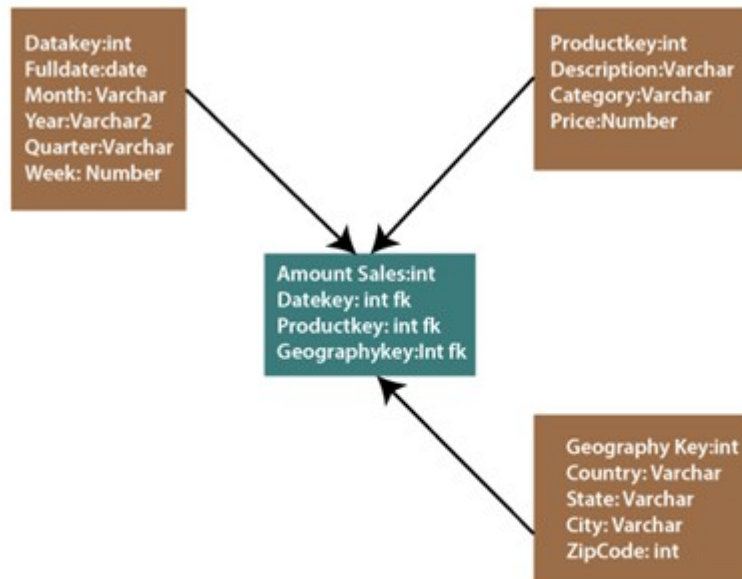
A generic data modeling life cycle



Example of Conceptual Data Model

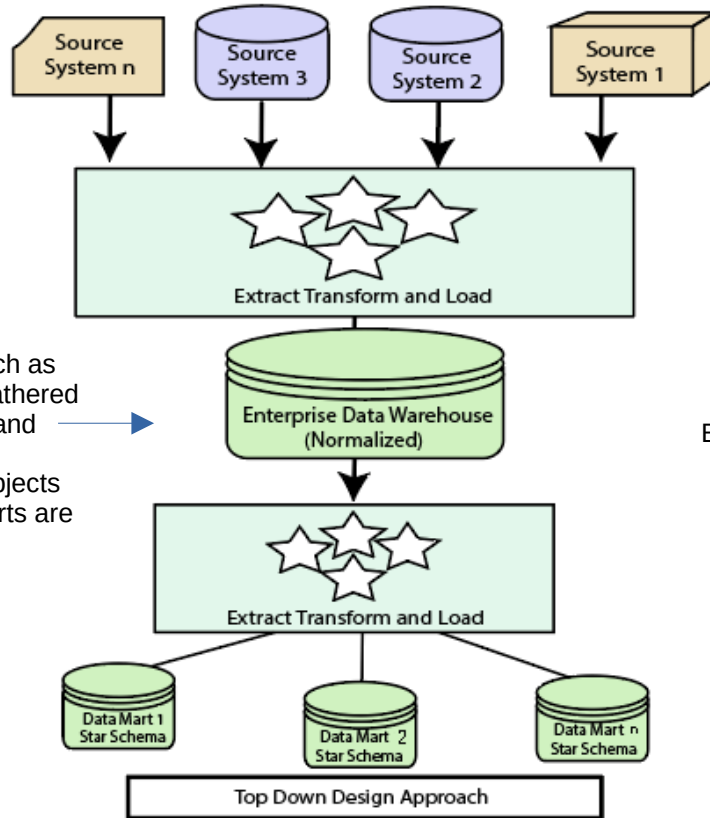


Example of Logical Data Model



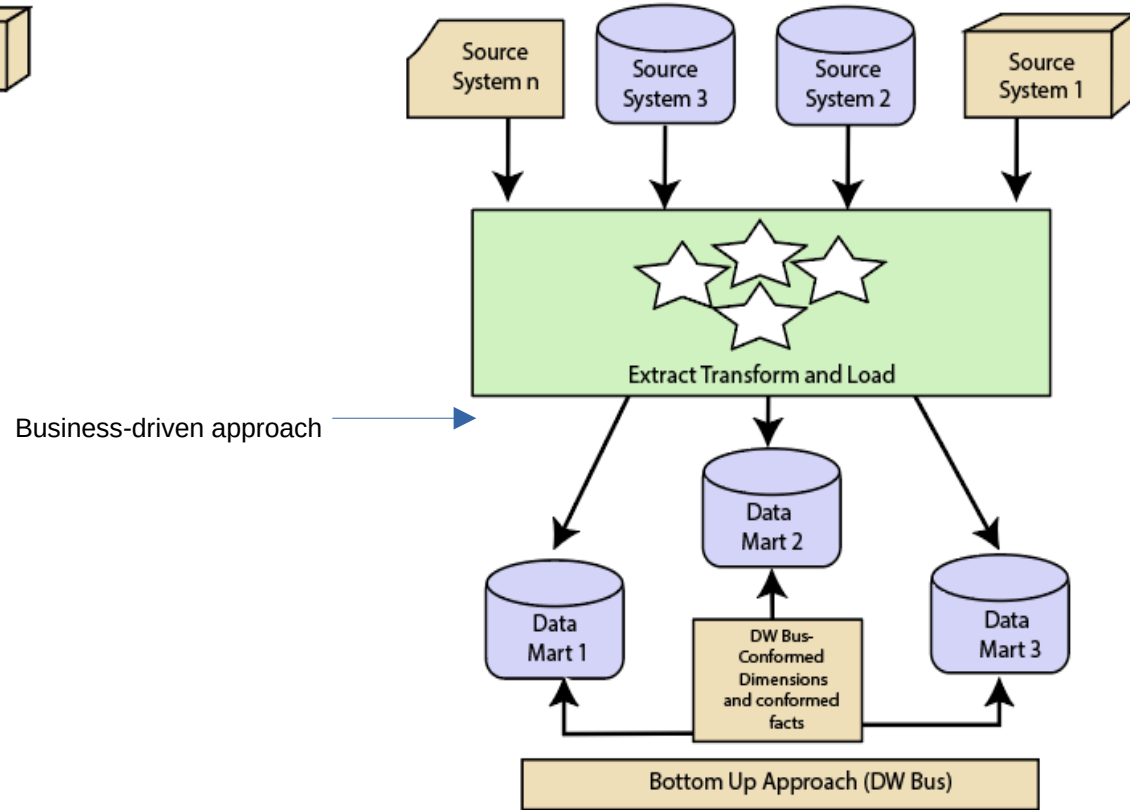
Example of Physical Data Model

Design Approach



Data-driven approach as the information is gathered and integrated first and then business requirements by subjects for building data marts are formulated.

Top Down Design Approach



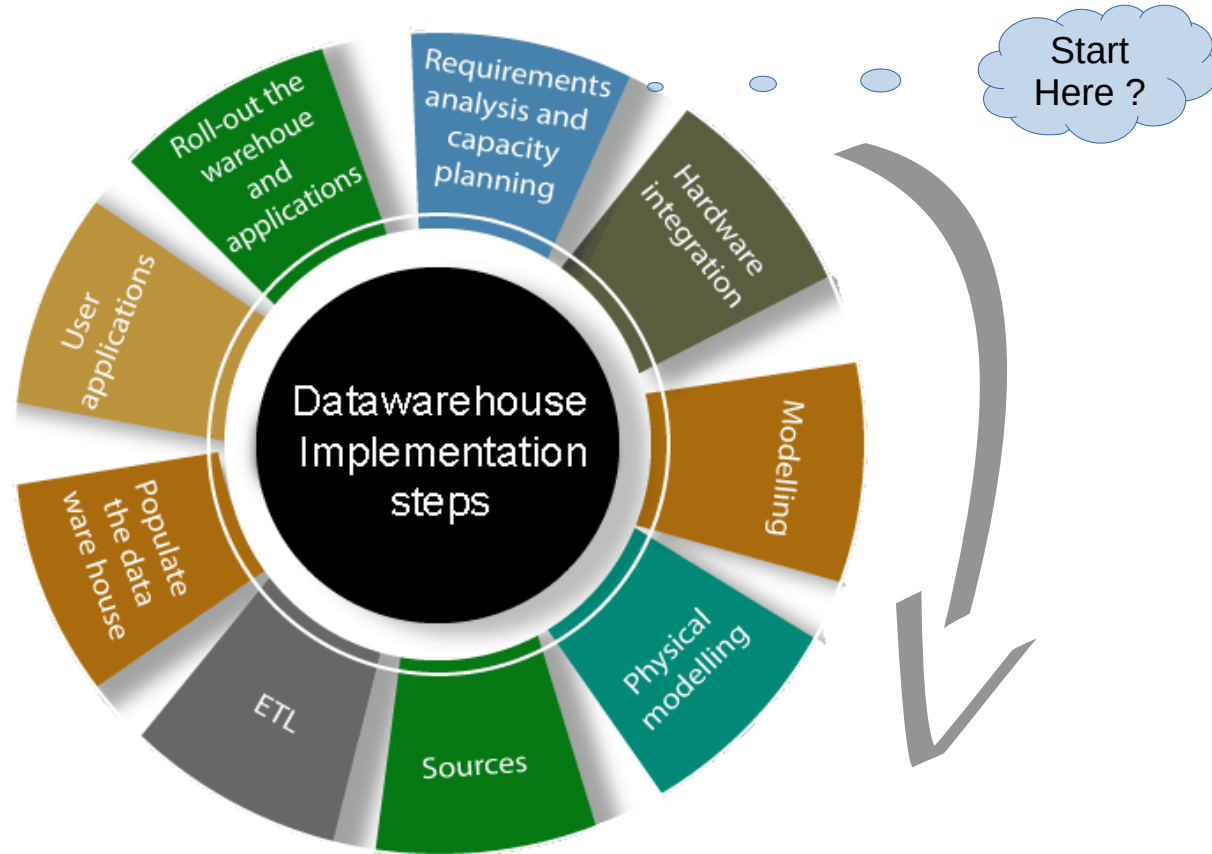
Business-driven approach

Bottom Up Design Approach

Comparison

Top-Down Design Approach	Bottom-Up Design Approach
Breaks the vast problem into smaller subproblems.	Solves the essential low-level problem and integrates them into a higher one.
Inherently architected- not a union of several data marts.	Inherently incremental; can schedule essential data marts first.
Single, central storage of information about the content.	Departmental information stored.
Centralized rules and control.	Departmental rules and control.
It includes redundant information.	Redundancy can be removed.
It may see quick results if implemented with repetitions.	Less risk of failure, favorable return on investment, and proof of techniques.

DW implementation ?



Data Warehousing Terms

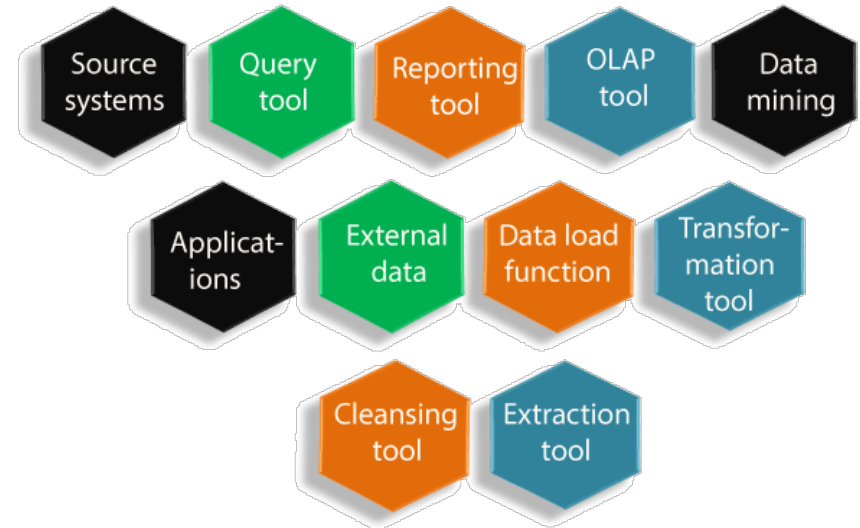
Metadata – Data about data.

E.g Indexes of Books, serving a date referring relevant data.

Metadata Repository -

- a) Business metadata – ownership, definition, policies
- b) Operational Metadata – data lineage – history etc.
- c) Mapping from OLTP to DW
- d) Algorithms for Summarization – dimensionality, granularity, aggregation, summarizing etc.

Data Warehouse Metadata



OPERATIONAL Data Stores

Subject-oriented, integrated, volatile, current valued data store, containing only detailed corporate data.

a) **Subject-oriented** - It is organized around the significant information subject of an enterprise.

E.g

In a university, the subjects may be students, lecturers and courses while in the company the subjects might be users, salespersons and products.

b) **Integrated** - it is a group of subject-oriented record from a variety of systems to provides an enterprise-wide view of the information.

c) **Current-valued** - an ODS is up-to-date and follow the current status of the data.

An ODS does not contain historical information. Since the OLTP system data is changing all the time, data from underlying sources refresh the ODS as generally and frequently as possible.

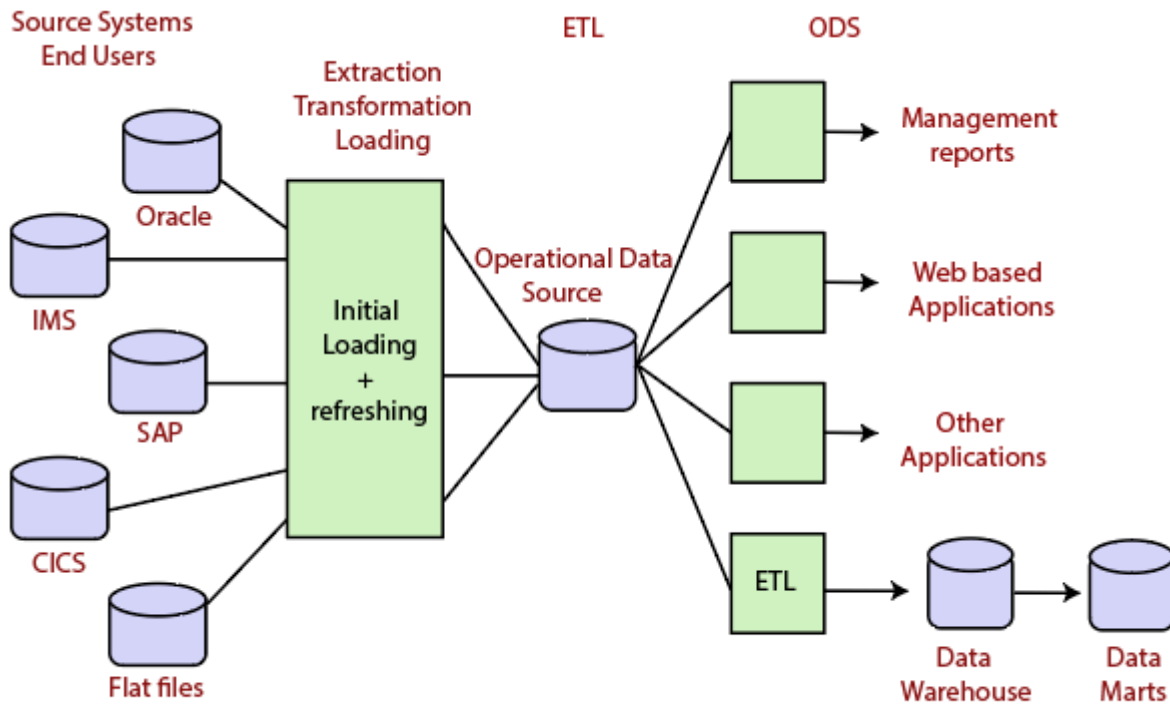
d) **Volatile** - That is, the data in the ODS frequently changes as new data refreshes the ODS.

e) **Detailed** - That is, ODS is detailed enough to serve the need of the operational management staff in the enterprise.

ODS Design and Implementation

Data is refreshed generally and frequently, suitable checks are required to ensure the quality of data after each refresh.

An ODS is a read-only database other than regular refreshing by the OLTP systems. Customer should not be allowed to update ODS information.



Operational Data Store Structure

Populating an ODS contains an acquisition phase of **extracting, transforming** and **loading** information from OLTP source systems.

This procedure is ETL. Completing populating the database, analyze for anomalies and testing for performance are essential before an ODS system can go online.

Delta – ODS vs DW

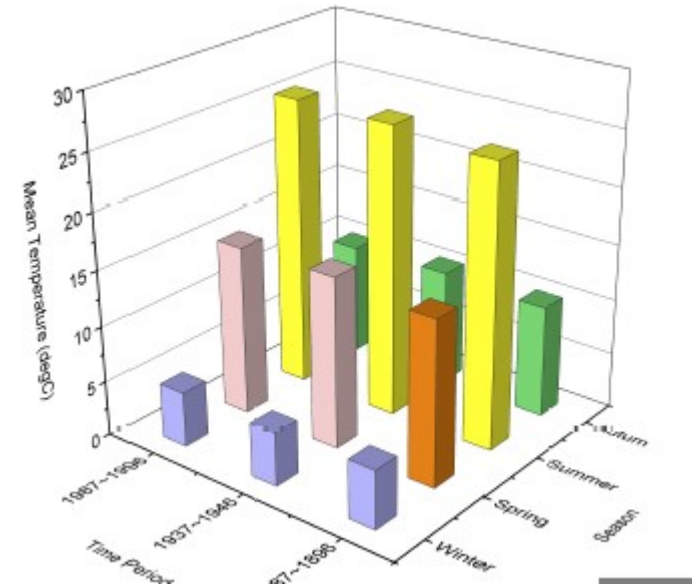
Operational Data Stores	Data Warehouse
ODS means for operational reporting and supports current or near real-time reporting requirements.	A data warehouse is intended for historical and trend analysis, usually reporting on a large volume of data.
An ODS consist of only a short window of data .	A data warehouse includes the entire history of data .
It is typically detailed data only.	It contains summarized and detailed data.
It is used for detailed decision making and operational reporting.	It is used for long term decision making and management reporting.
It is used at the operational level.	It is used at the managerial level.
It serves as conduct for data between operational and analytics system.	It serves as a repository for cleansed and consolidated data sets.
It is updated often as the transactions system generates new data.	It is usually updated in batch processing mode on a set schedule.

Data cube

Multiple dimensions of data representation

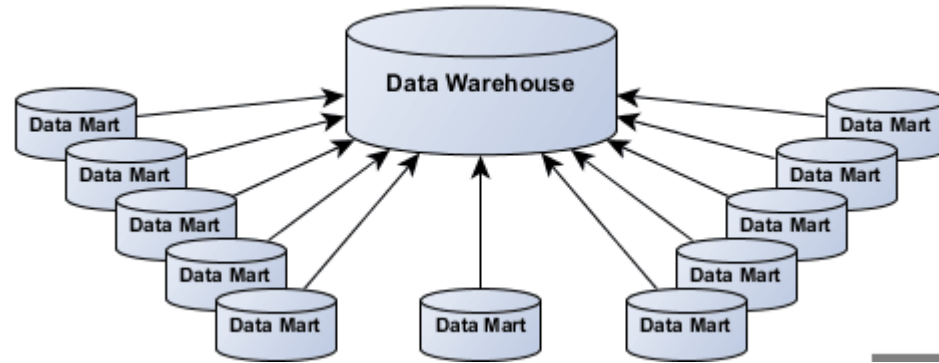
A	B	C	D	E	F	G	H	I
Sales Data (Pics) of 2022								
Salesman	Jan	Feb	Mar	Apr	May	Jun	Total Sales	
Marry	70	80	75	60	72	55	↑ 412	
Joe	30	48	35	45	25	37	↓ 220	
Bob	65	54	49	54	35	65	→ 322	
Taylor	85	71	68	77	88	73	↑ 462	
James	55	25	45	50	53	30	↓ 258	
Richard	35	45	15	45	45	25	↓ 210	
Tessy	75	66	59	65	56	30	→ 351	
Thompson	29	35	45	48	35	55	↓ 247	

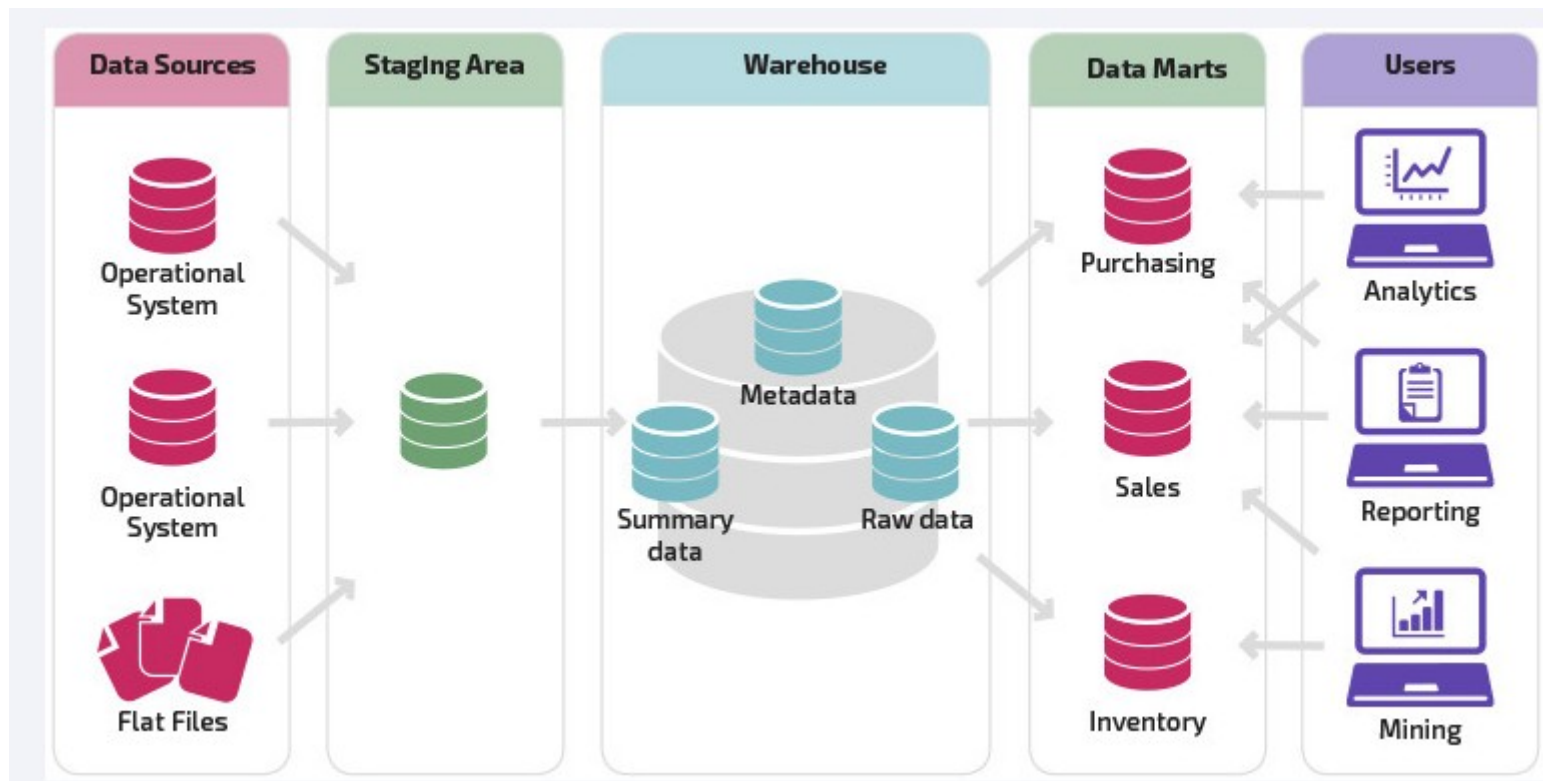
3D view



Data Mart

Subset of organizational data important, valuable to different processes or departments of organization.

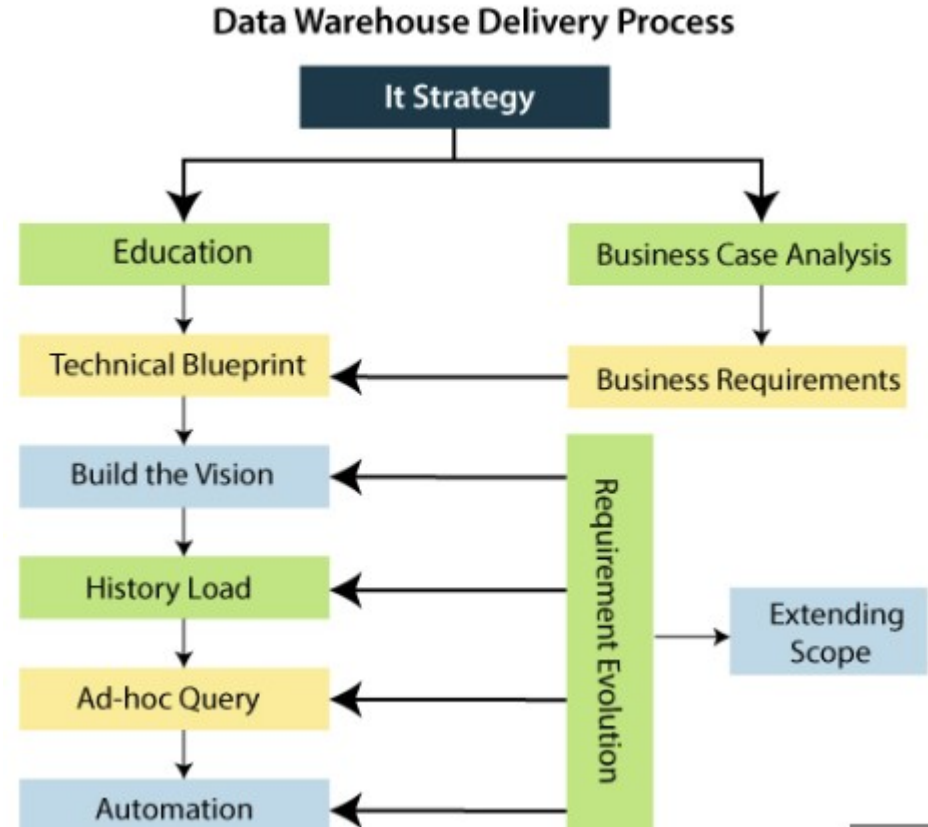




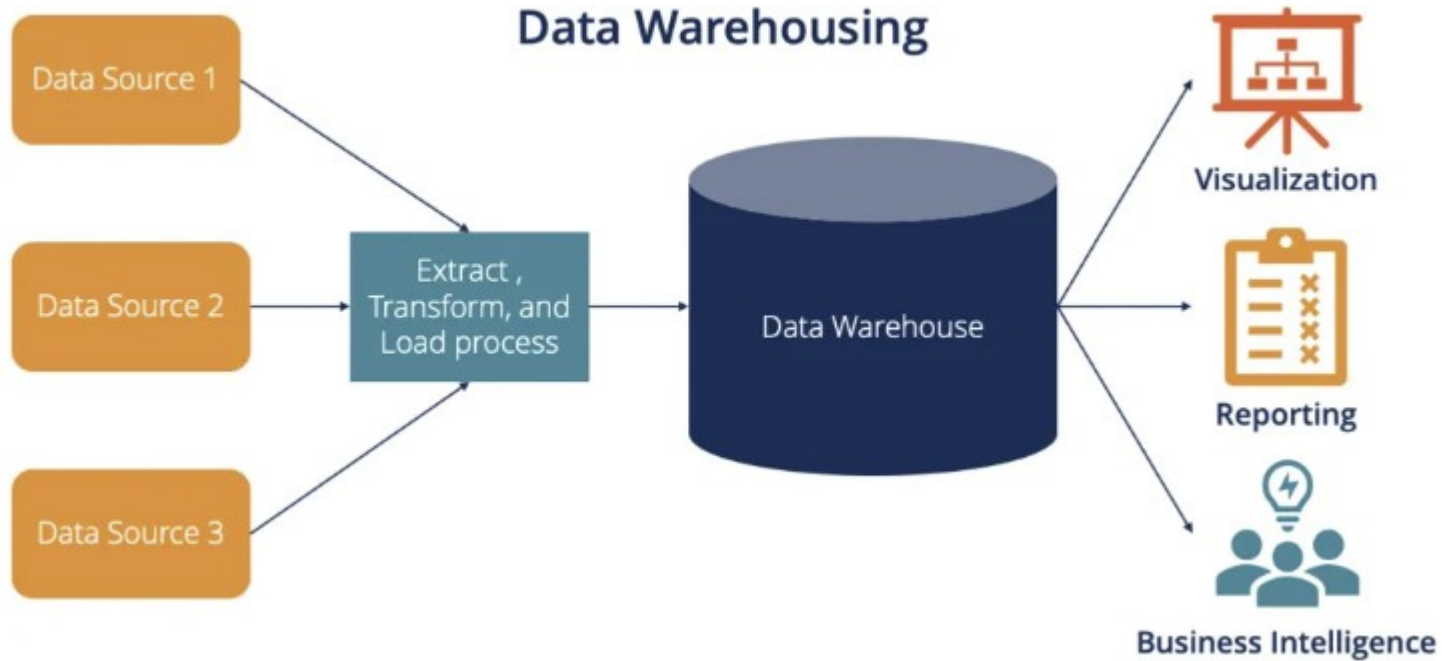
Data Warehouse – Delivery process

Data warehouse must be flexible leading to future growth as business expands, grows.

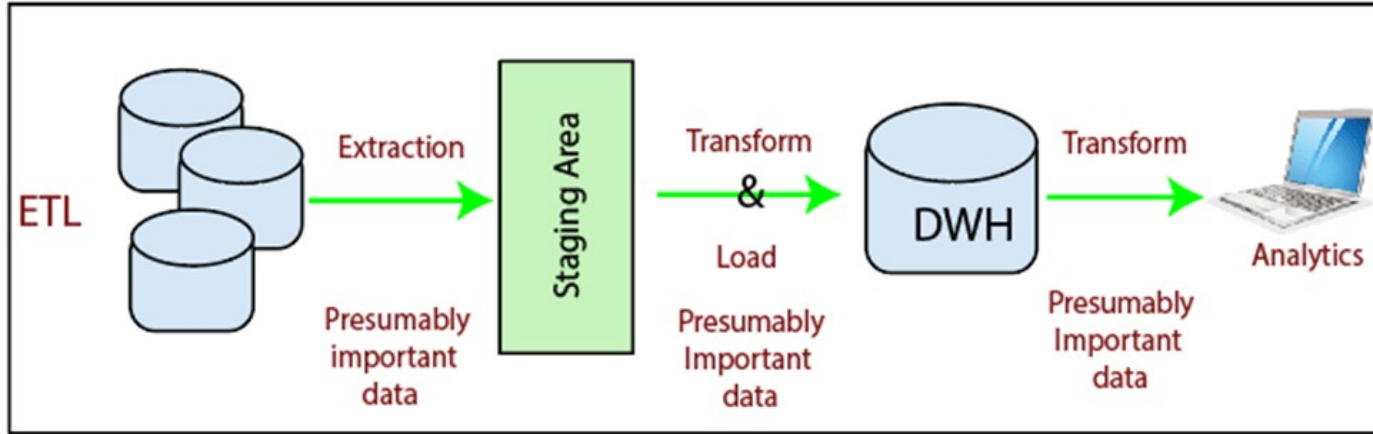
Delivery process is broken into phases to reduce the project and delivery risk.



Data Warehousing – System Processes

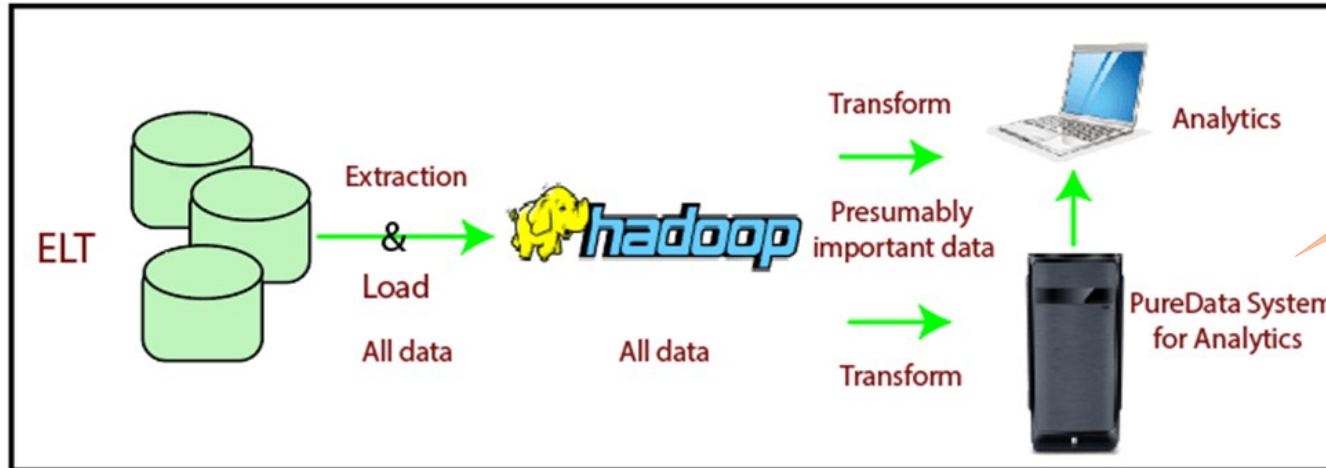
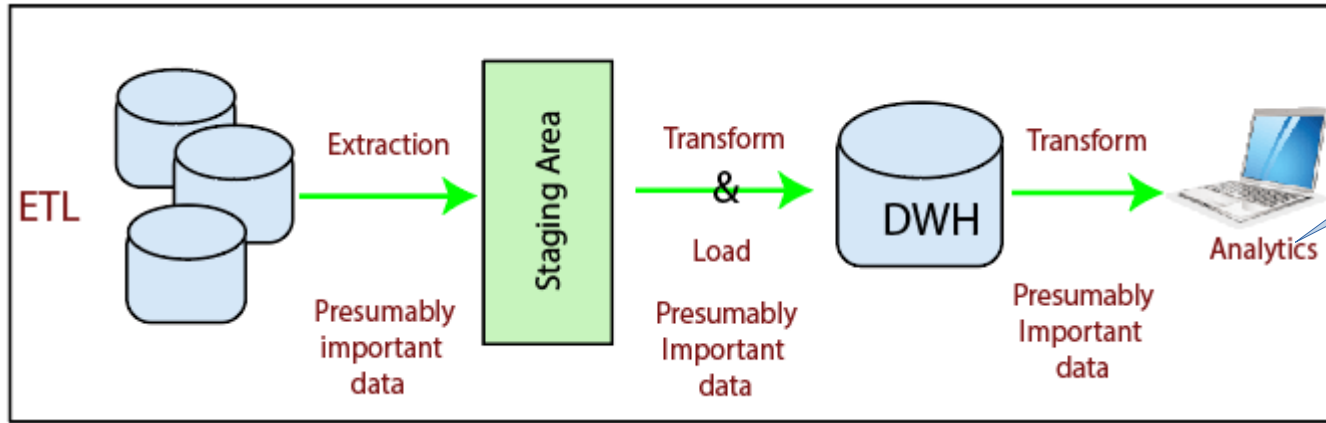


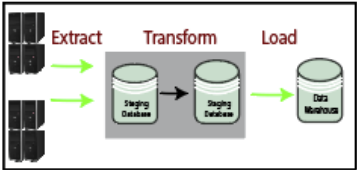
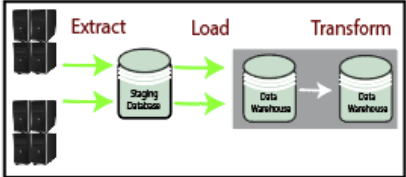
ETL (Extract, Transform, and Load) Process



Mechanism of extracting information from source systems and bringing it into the data warehouse is commonly called ETL, which stands for **Extraction**, **Transformation** and **Loading**.

Difference between ETL and ELT



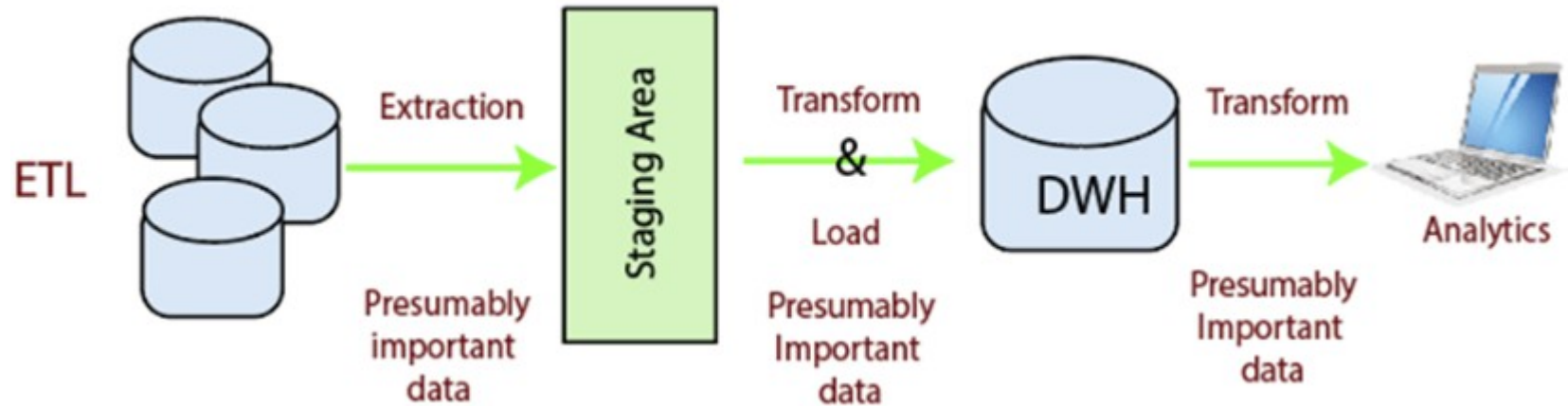
Basics	ETL	ELT
Process	Data is transferred to the ETL server and moved back to DB. High network bandwidth required.	Data remains in the DB except for cross Database loads (e.g. source to object).
Transformation	Transformations are performed in ETL Server.	Transformations are performed (in the source or) in the target.
Code Usage	Typically used for <ul style="list-style-type: none"> Source to target transfer Compute-intensive Transformations Small amount of data 	Typically used for <ul style="list-style-type: none"> High amounts of data
Time-Maintenance	It needs high maintenance as you need to select data to load and transform.	Low maintenance as data is always available.
Calculations	Overwrites existing column or Need to append the dataset and push to the target platform.	Easily add the calculated column to the existing table.
Analysis	 <pre> graph LR Source[Source] -- Extract --> SD1[Staging Database] SD1 -- Transform --> SD2[Staging Database] SD2 -- Load --> DW[Data Warehouse] </pre>	 <pre> graph LR Source[Source] -- Extract --> SD[Staging Database] SD -- Load --> DW1[Data Warehouse] DW1 -- Transform --> DW2[Data Warehouse] </pre>

Extract and Load Process

Data extraction takes data from the source systems.

Data load takes the extracted data and loads it into the data warehouse.

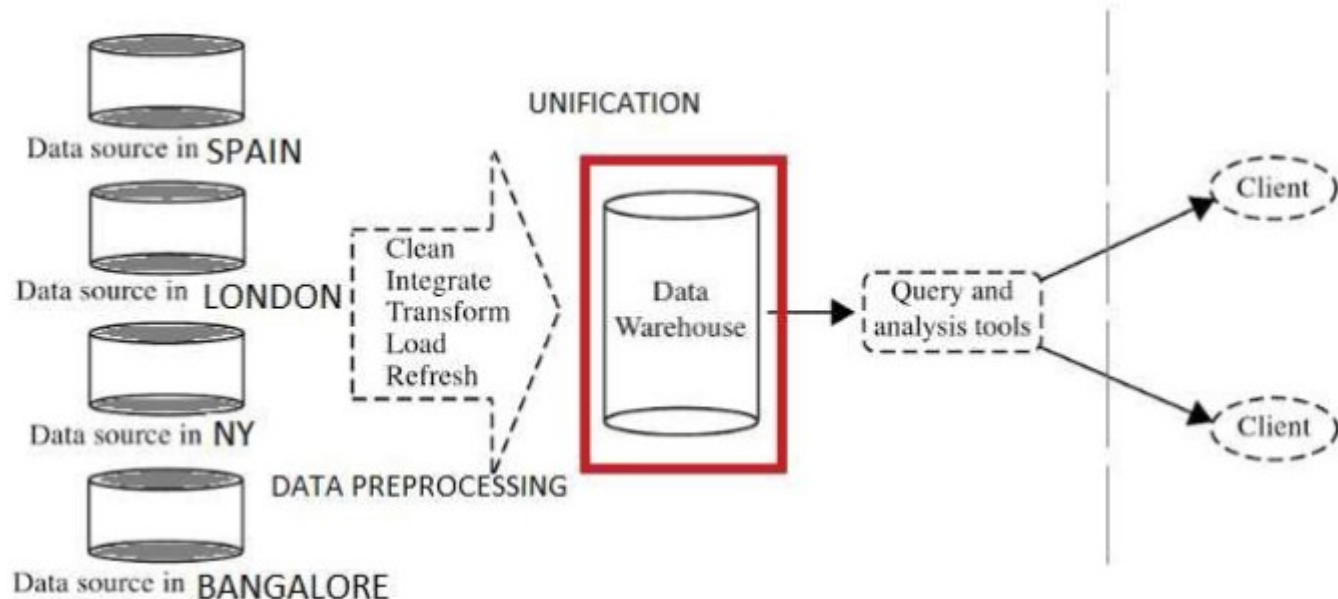
Data before ETL must be reconstructed which is received from Data sources.



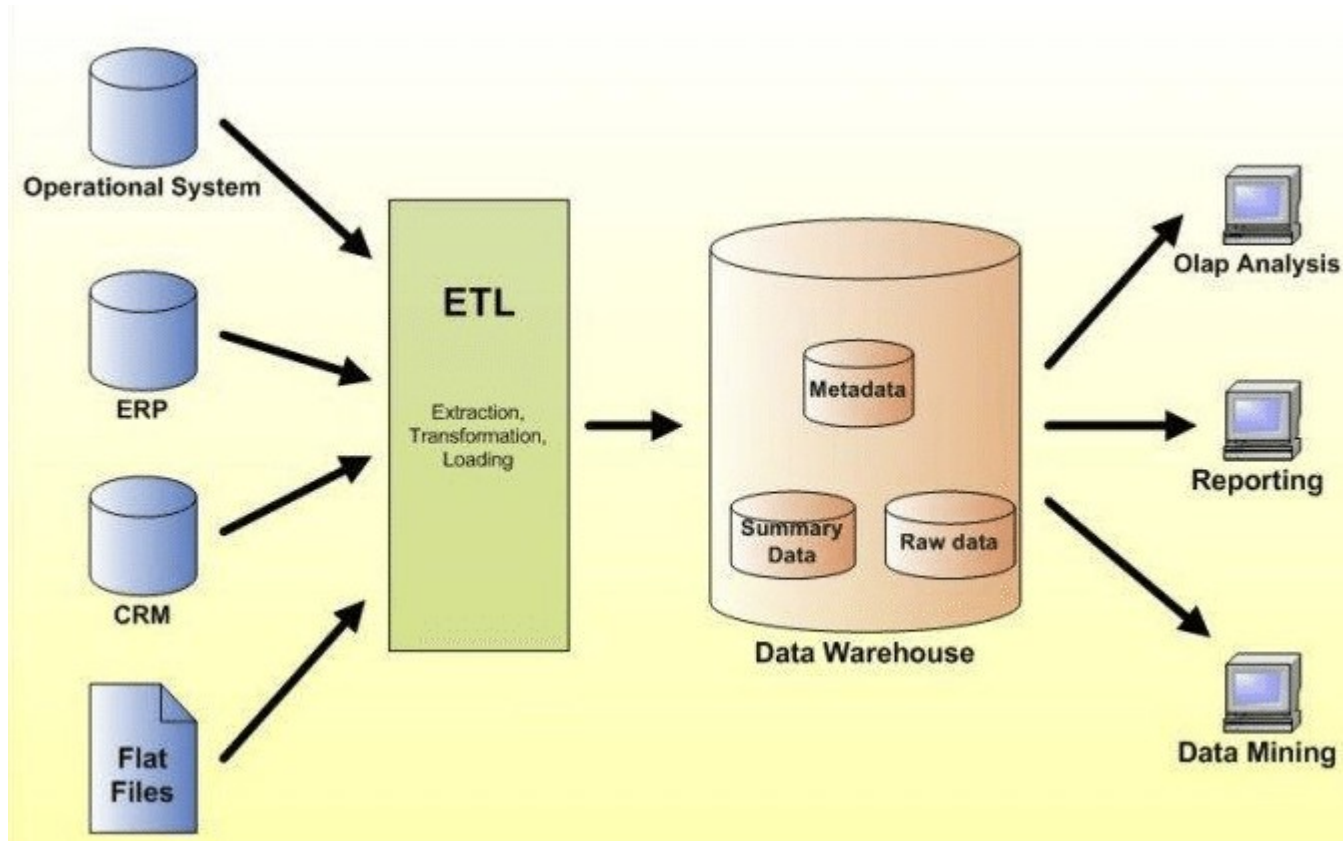
Clean & Transform

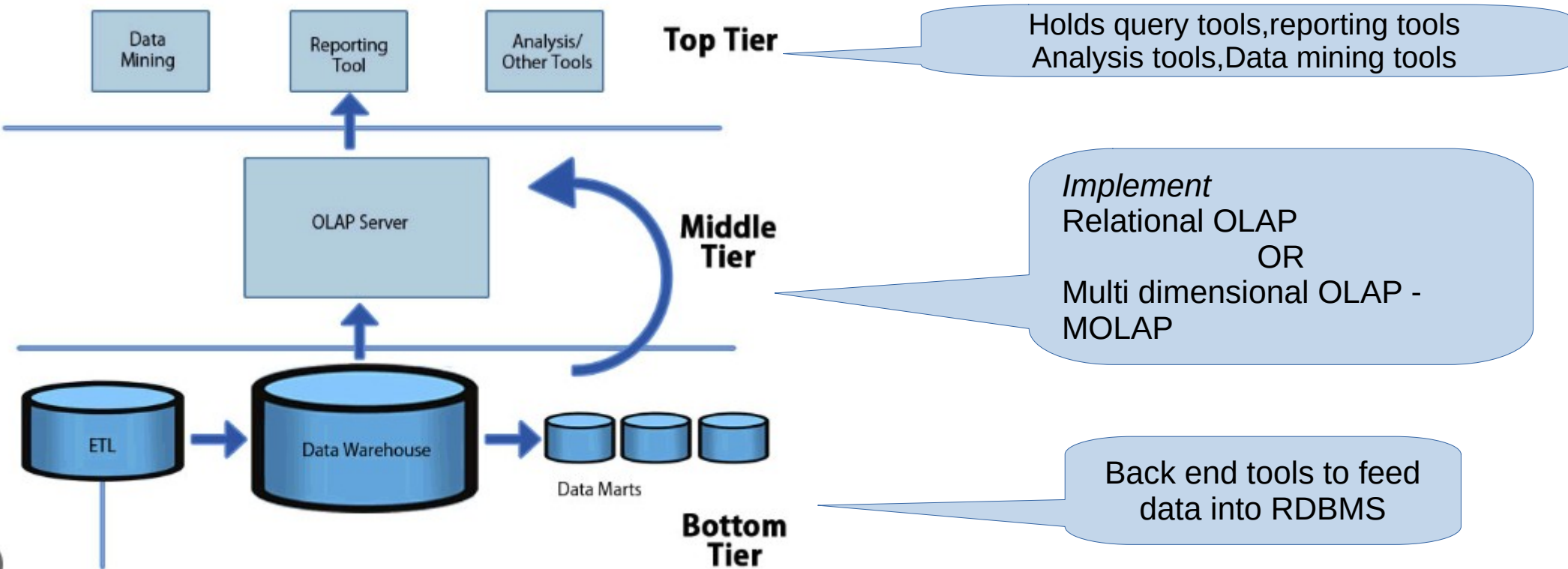
After data is extracted and loaded into the temp data store -

- a) Clean and transform the loaded data into a structure.
- b) Partition the data.
- c) Aggregation – to speed up common queries.



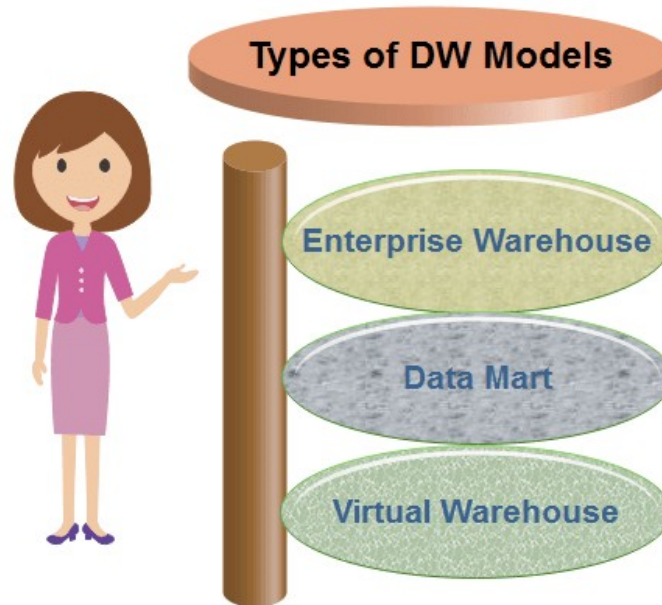
Data warehousing Architecture



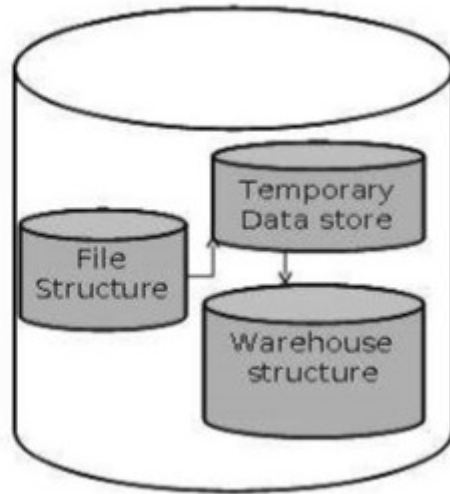
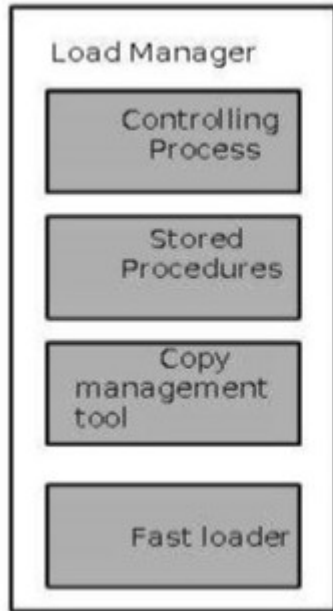


Data Warehouse Models

- Virtual Warehouse** – View created over an Operational data warehouse
- Data Mart** – a subset of organization wide data.
- Enterprise Warehouse** – Collects all information spanning subjects of entire organization



Load Manager Arch

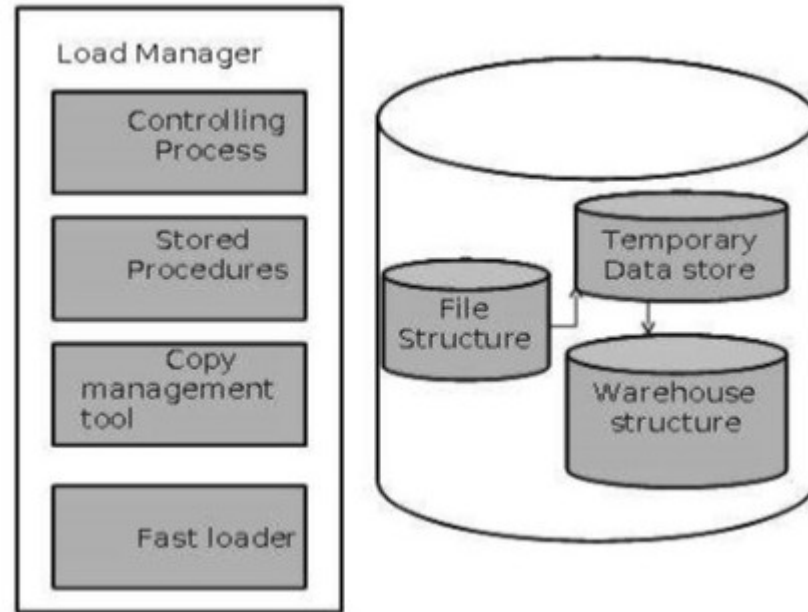


1) Extract data from source systems.

2) Faster loading into temp data store.

3) Perform transformations – removing not required columns etc.

Warehouse Manager Arch



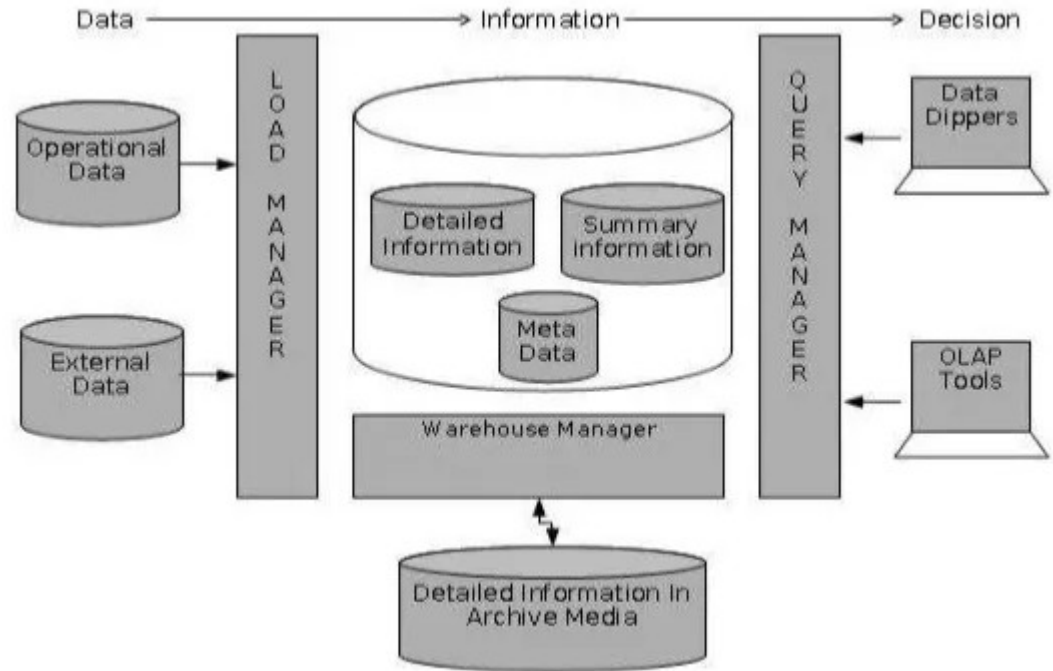
Ware House manager operations

- 1) Analyzes data to perform consistency and referential integrity checks.
- 2) Creates Indexes, business views, partitions views against base data.
- 3) Transforms, merges, the source data into published data warehouse.
- 4) Backup the data in the data warehouse.
- 5) Archives the data.

Query Manager Arch

Responsible for

- 1) Directing queries to the suitable tables.
- 2) Directing queries to appropriate tables, performance increase by decreasing latency.
- 3) Scheduling execution of queries from user.



Data warehousing - OLAP

Types of OLAP Servers

1) Relational – ROLAP

2) Multi dimensional – MOLAP

3) Hybrid – HOLAP

4) Specialized SQL Servers

OLAP – OnLine Analytical Processing

ROLAP – Relational OnLine Analytical Processing

MOLAP – Multidimensional OnLine Analytical Processing

HOLAP – Hybrid OnLine Analytical Processing

DOLAP – **Desktop**/Database OnLine Analytical Processing

WOLAP – **Web Enabled** OnLine Analytical Processing

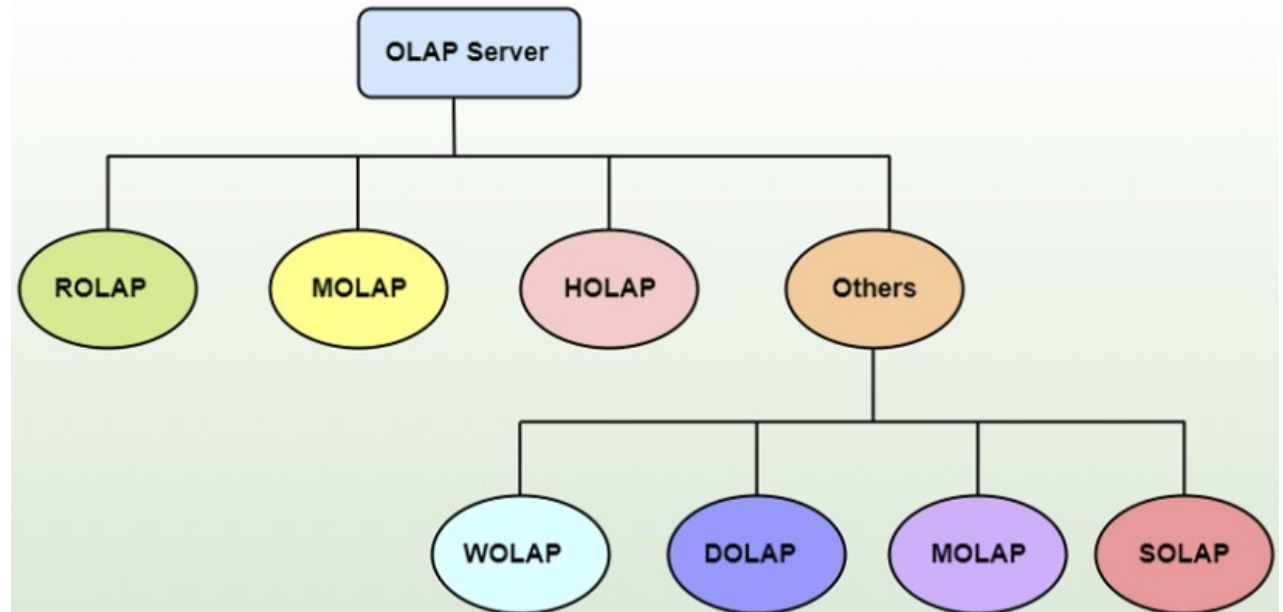
SOLAP -- Spatial On-line Analytical Processing

Relational OLAP

It's implementation of aggregation navigation logic.

Optimizing for each DBMS back end.

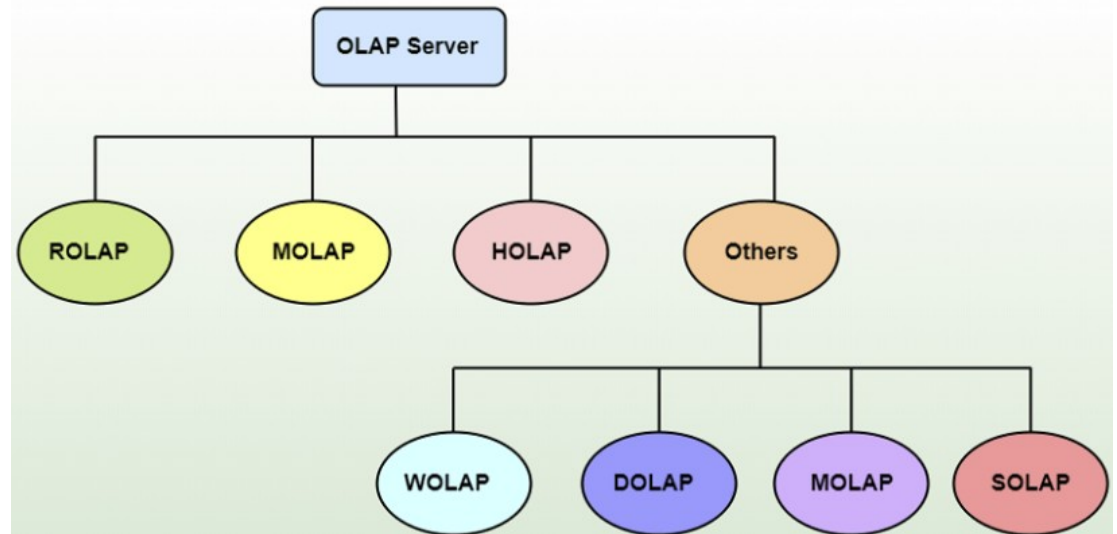
Additional tools and Services



Multidimensional OLAP

Array based Multidimensional storage engines for multidimensional views of data.

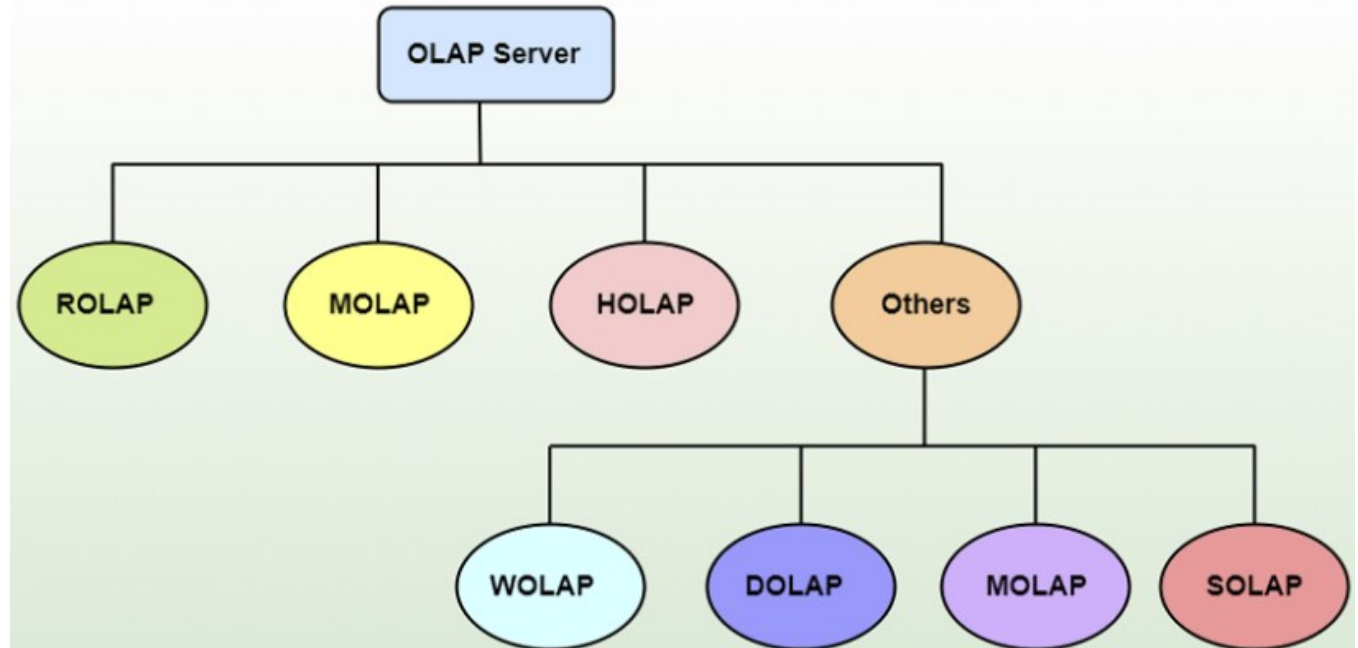
Due to multi-dimensionality – data storage may be sparse or densely populated – MOLAP then uses 2 levels data storage representation to handle data.



Hybrid HOLAP

Combination of **ROLAP** and **MOLAP**.

Provides Scalability, fast computation



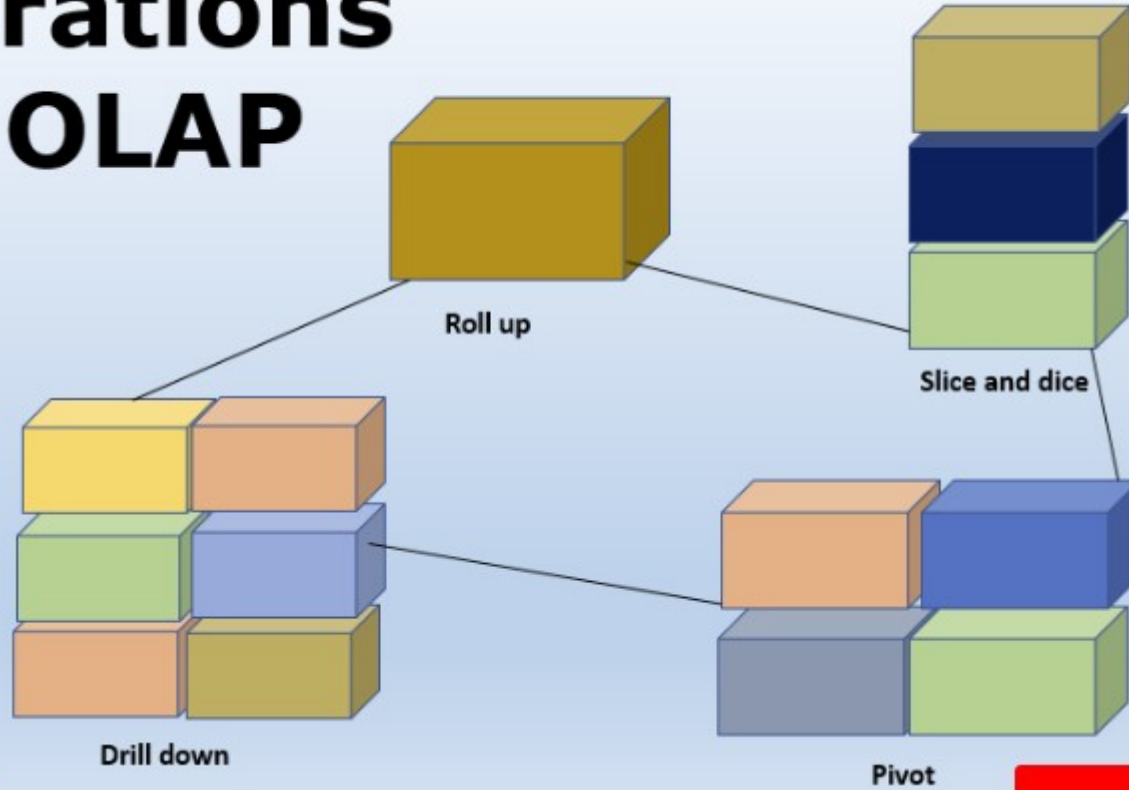
Specialized SQL Servers

Provides for

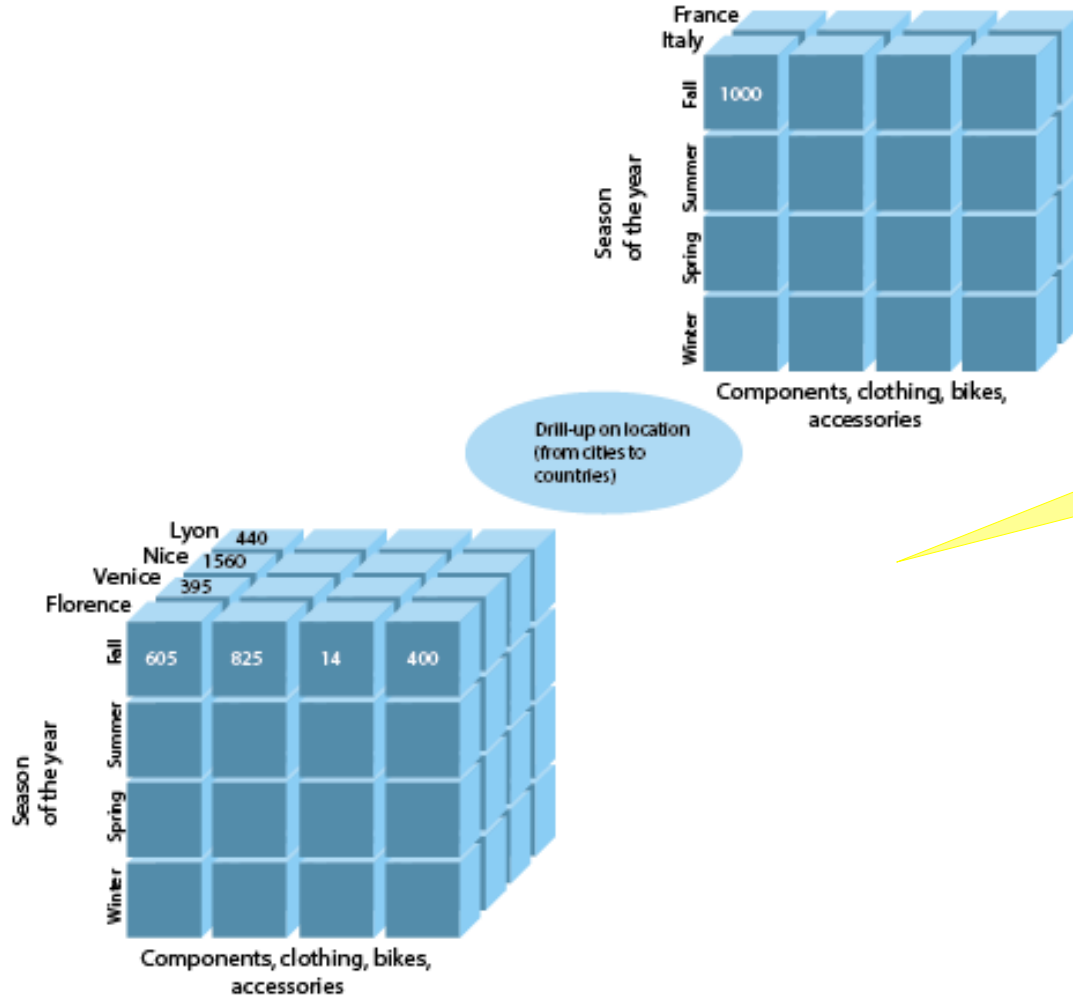
- a) Advanced query language
- b) Query processing support.

OLAP operations

Operations in OLAP



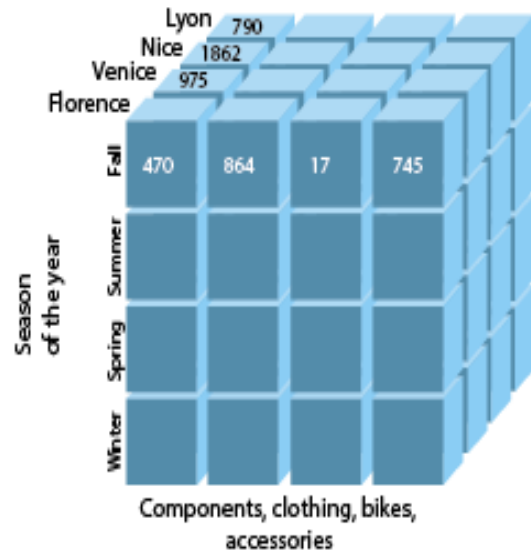
Roll up or Drill up



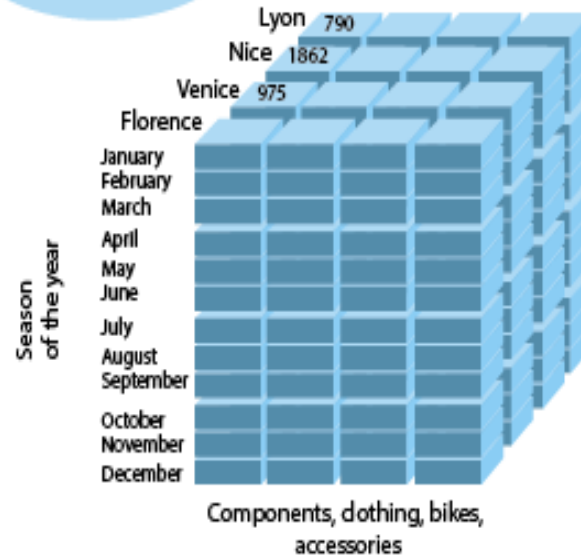
Performs aggregation on data cube
1) By Climbing up a concept hierarchy for a dimension

2) By dimension reduction

Drill Down



Drill down on time (from quarters to month)

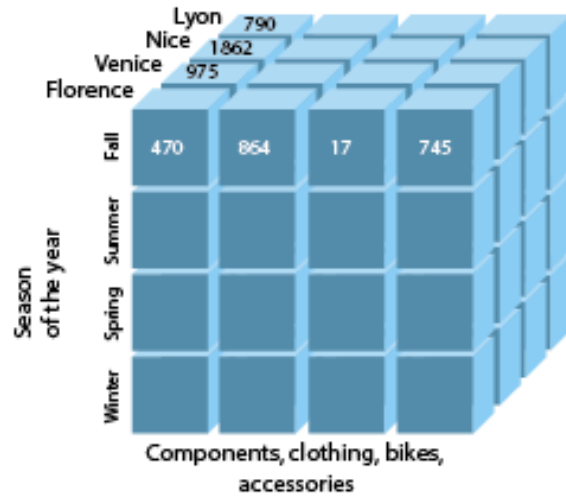


Performs

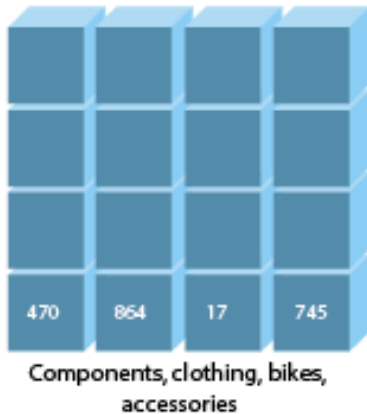
1) By stepping down a concept hierarchy for a dimension.

2) By Introducing a new dimension.

Slice

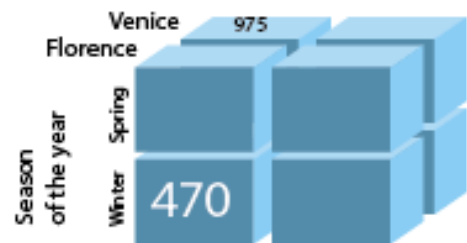


Slice
for time
="winter"



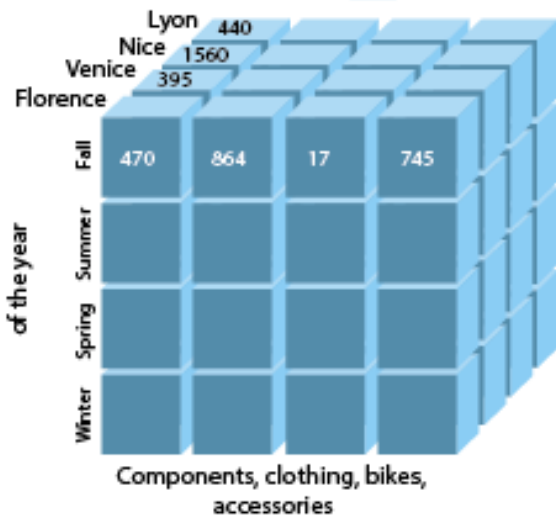
Selects one particular dimension from a given data cube and outputs new data cube

Dice

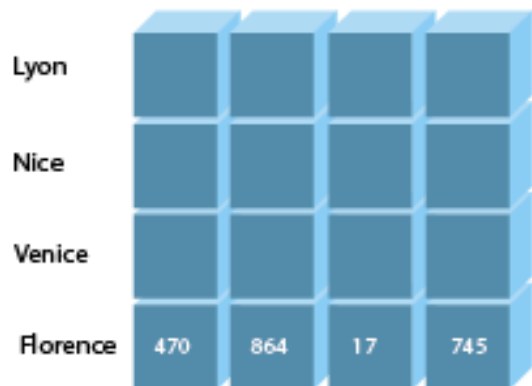


Dice for (location = "Venice" or
"Florence")
and (season = "Winter" or "Spring") and
(item = "components" or "clothing")

Selects two or more dimensions from data cube and provides a new data cube

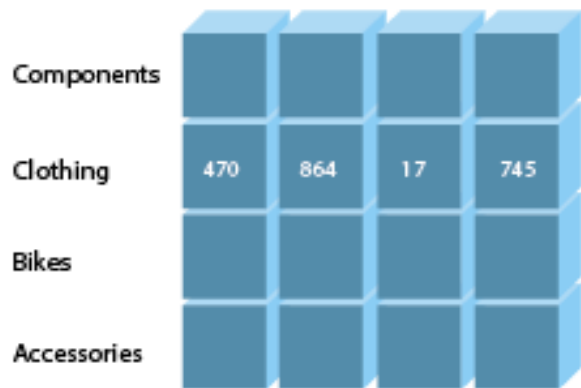


Pivot



Components, clothing, bikes,
accessories

Slice
for season
="winter"

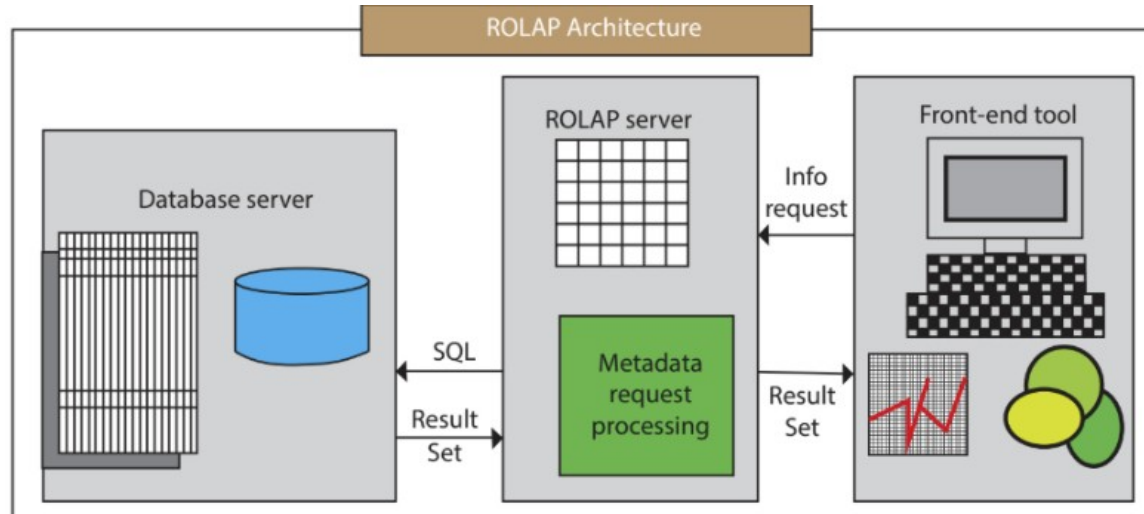


Components, clothing, bikes,
accessories

Also kind of Rotation.

Rotates the data axes in view in order to
provide alternative presentation of data

Data Warehousing - ROLAP



ROLAP

Highly Scalable

Tools can analyze large volumes of data across multiple dimensions.

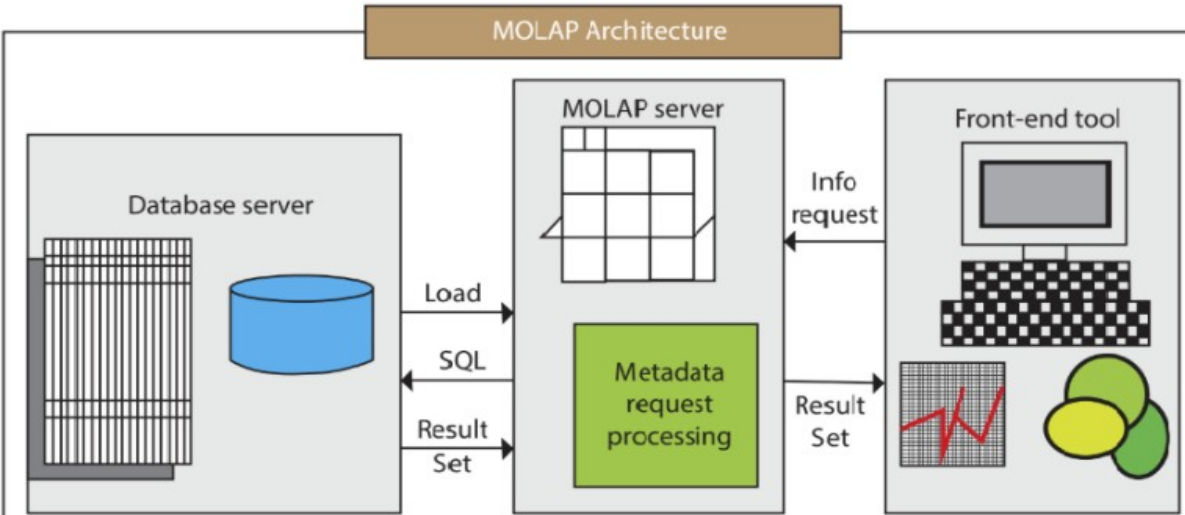
Tool can store and analyze highly volatile and changeable data

Limitations

SQL functionality is constrained, slow performance.

Difficulty in aggregated tables update

MOLAP



Allows fastest indexing, retrieval.

Slicing, dicing operations are fast

Complex calculations are fast.

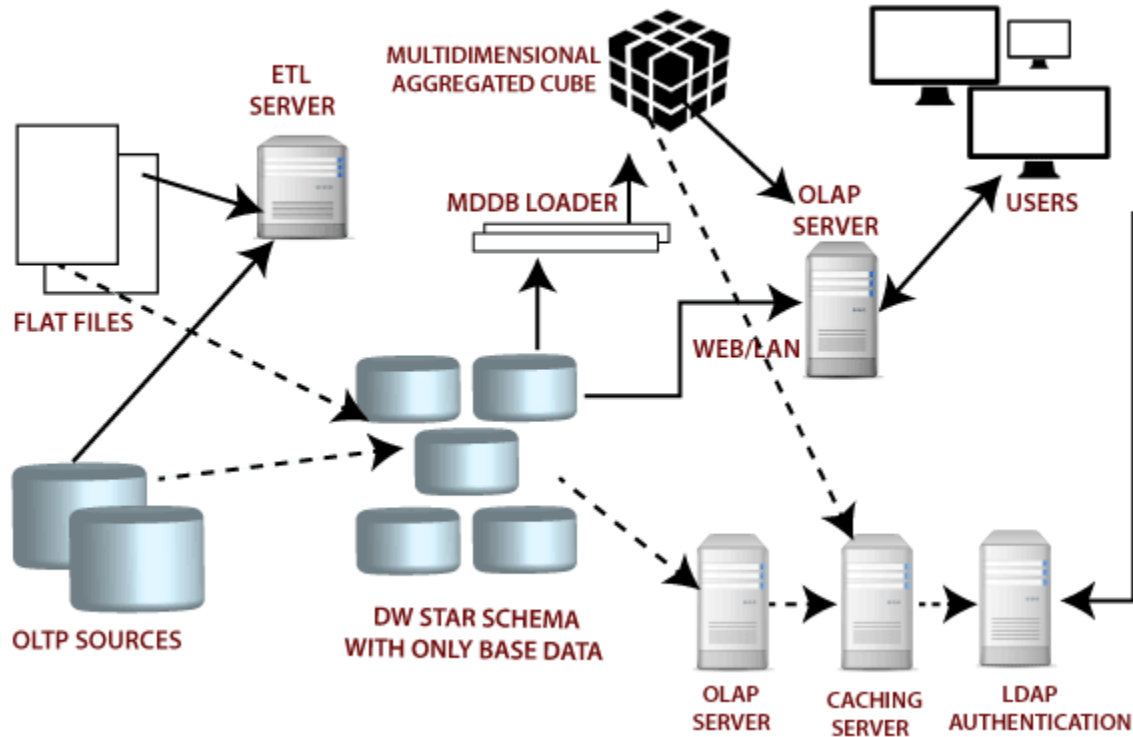
Limited in the amount of information it can handle

Incapable to handle detailed data

Storage utilization may be low if data set is sparse

HOLAP

HOLAP Architecture



It provides fast access at all levels of aggregation.

HOLAP balances the disk space requirement, as it only stores the aggregate information on the OLAP server and the detail record remains in the relational database.

HOLAP architecture is very complicated because it supports both MOLAP and ROLAP servers.

Schema

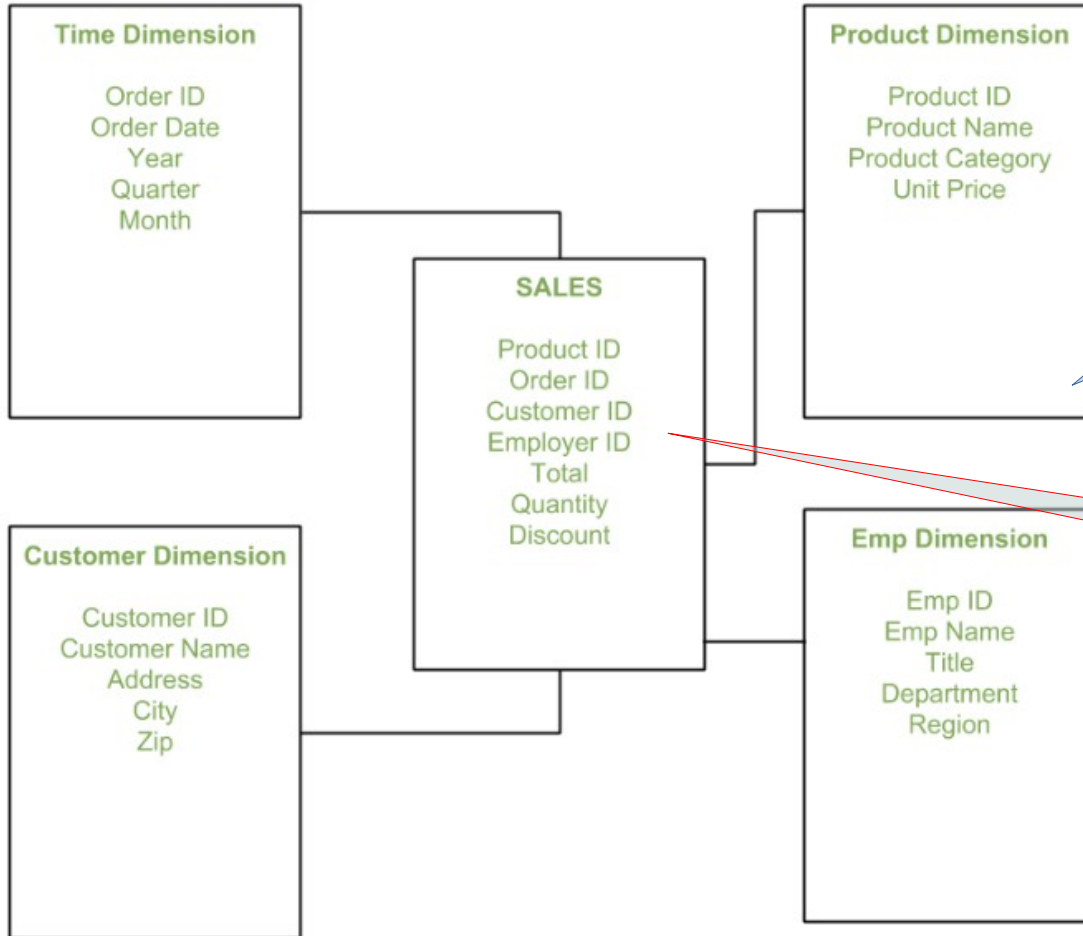
Schema is logical description of entire database.

Includes name and description of all records types including all associated data items and aggregates.

The schema in data warehouse could be -

- 1) Star
- 2) Snowflake
- 3) Fast Constellation

Star Schema



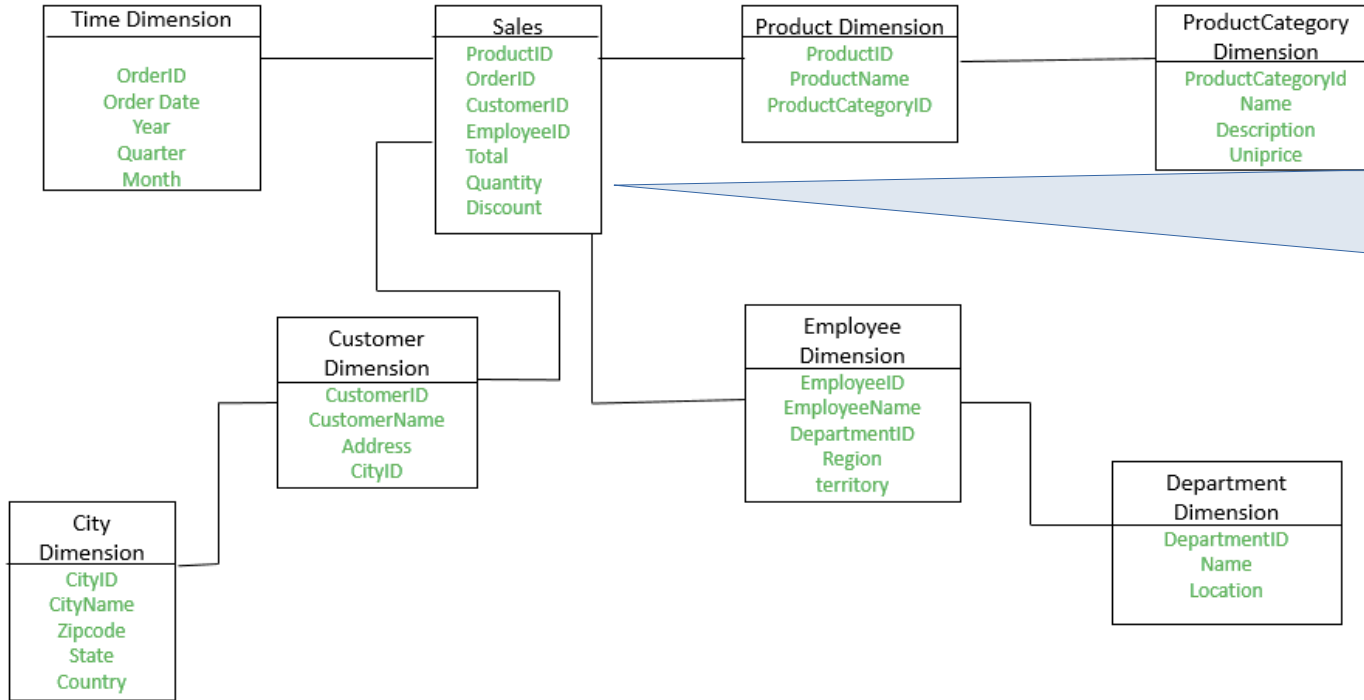
Representation in

1) One dimensional table only.

2) Dimension table contains set of attributes.

Fact table contains keys to each of four dimensions.

Snowflake Schema



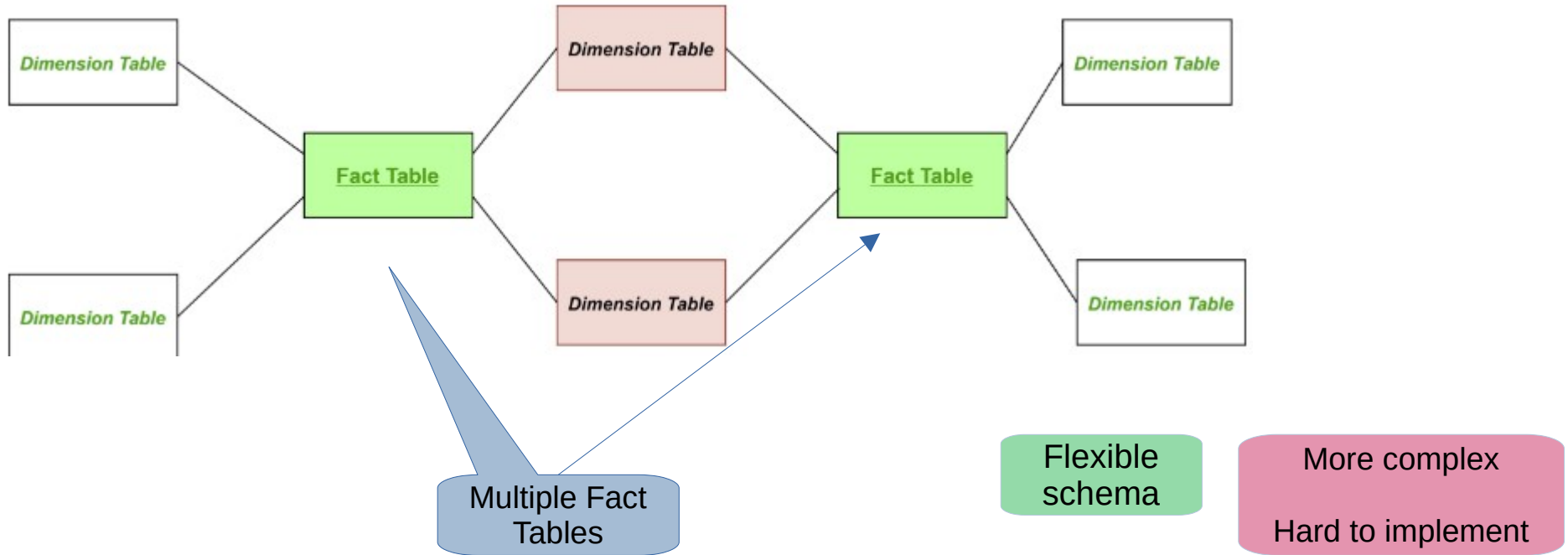
Centralized fact table is connected to multiple dimensions.

Dimensions are present in a normalized form in multiple related tables.

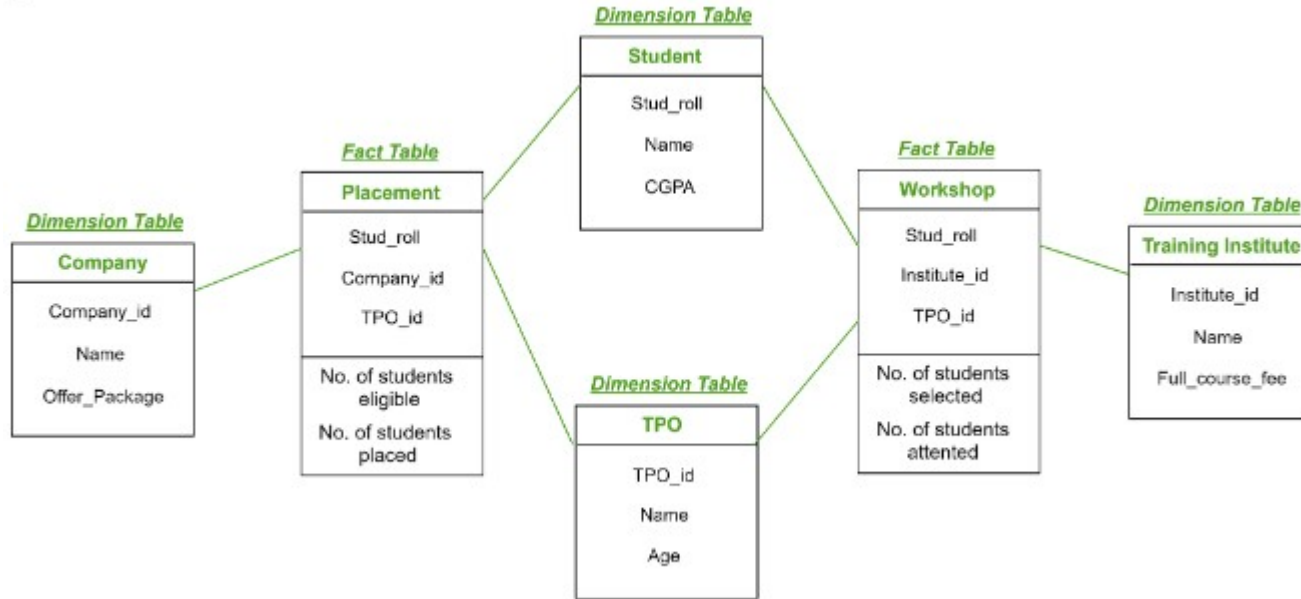
snowflake schema uses small disk space.

There are multiple tables, so performance is reduced.

Fact constellation or Galaxy



Eg



Data Warehousing - Partitioning Strategy

Horizontal
Partitioning

Partitioning function

Partitioning schema

Big table

Section

May 2019

File group
May 2019

Section

June 2019

File group
June 2019

Section

July 2019

File group
July 2019

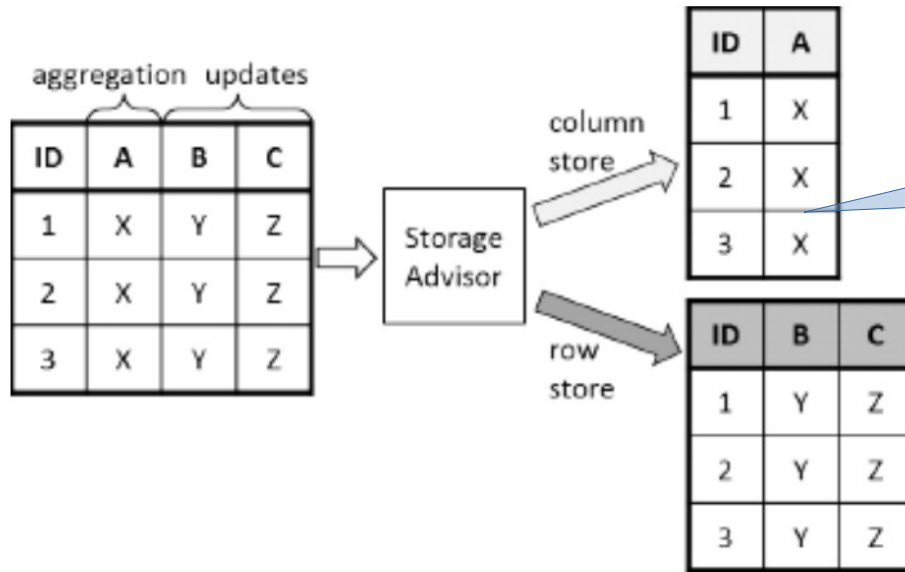
Enhance performance

Easy data management

Balancing various aspects of data

Partitioning by Time.
Partitioning by Different sized.
Segments.
Partitioning by Size of table.
Partitioning by Different dimension.
Round Robin Partitions.

Vertical Partitioning



Splits the data vertically

Can be done with
1) Normalization
2) Row Splitting

Finally

