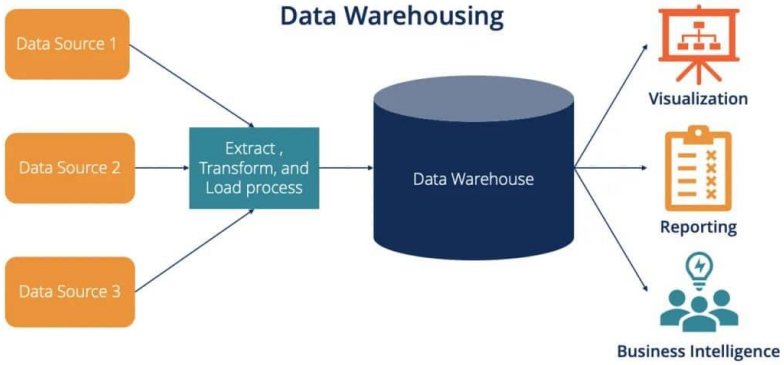
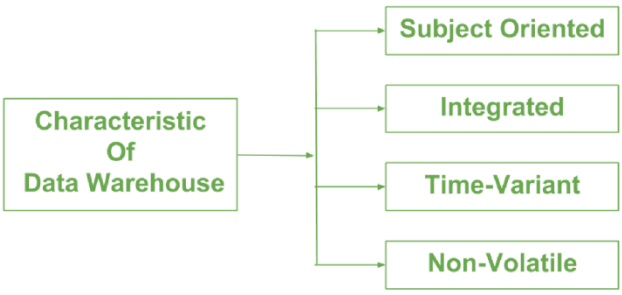
[Data Warehouse]

What is a data warehouse?

A data warehouse is a central repository of information that can be analyzed to make reports and informed decisions. A Data warehouse is typically used to connect and analyze business data from heterogeneous sources. The data warehouse is the core of the BI system which is built for data analysis and reporting.





What are the characteristics of a data warehouse?

- Subject-oriented: DW typically provides information on a topic (Sales, HR, Supply chain)
- Time-variant: Time variant keys (e.g., for the date, month, time) are typically present.
- Integrated: A data warehouse combines data from various sources. These may include a cloud, relational databases, flat files, structured and semi-structured data.
- Non-volatile: Prior data isn't deleted when new data is added. Historical data is preserved for comparisons, trends, and analytics

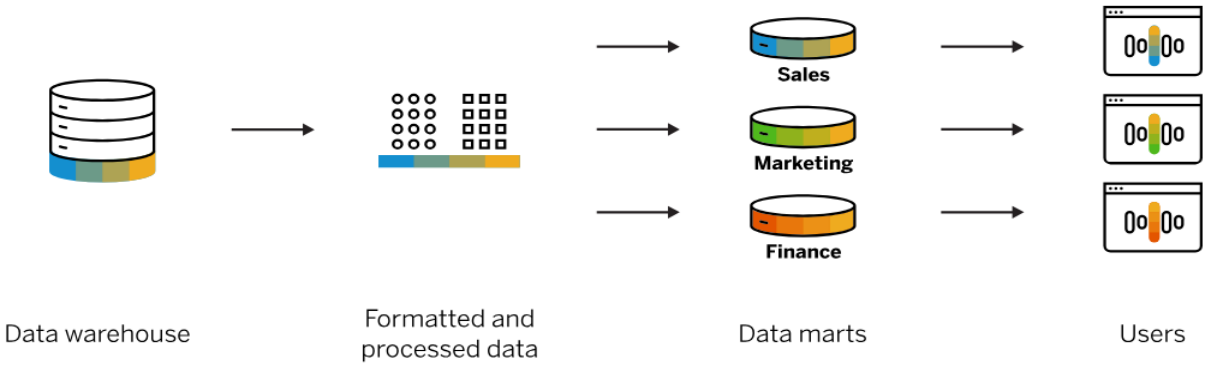
What is a database vs. a data warehouse?

Database	Data warehouse
Database stores the current data required to power an application.	Data warehouse stores current and historical for the purpose of analyzing the data.
Highly normalized, static schemas	Denormalized schemas, such as the Star schema or Snowflake schema

What is a bigdata vs. a data warehouse?

Big data	Data warehouse
Big data is a technology to store and manage large amounts of data.	Data warehouse is an architecture used to organize data from heterogenous sources
It takes structured. non-structured or semi structured data as input	It only takes structured data as an input
Big data doesn't follow any SQL queries to fetch data from database	In data warehouse we use SQL queries to fetch data from relational database

Data warehouse vs Data mart?



Data mart

- Decentralized, specific subject area
- A single community or department
- A single or a few sources, or a portion of the data warehouse
- Small, generally up to 10's of gigabytes

Data warehouse

- Centralized, multiple subject areas
- Organization-wide
- Many sources
- Large, can be 100's of gigabytes to petabytes

Data warehouse vs Data Lake?

Data lake

- Takes Structured, semistructured and unstructured data, from sensors, apps, websites, etc.
- May not have a predefined purpose, typically used for machine learning & deep analysis
- Used by Data engineers & Data Scientists
- Schema-on-Read

Data warehouse

- Takes Structured, processed data, from operational databases, applications and transactional systems.
- Predefined purposes for business intelligence batch reporting and visualization
- Used by Data engineers, business analysts and data analysts
- Schema-on-Write

What is the difference between OLTP and OLAP?

OLTP stands for Online Transaction Processing.
OLTP has the work to administer day-to-day transactions in any organization. The main goal of OLTP is data processing not data analysis.

OLAP stands for Online Analytical Processing.
OLAP systems have the capability to analyze database information of multiple systems at the current time. The primary goal of OLAP Service is data analysis and not data processing.

OLTP (RDBMS)	OLAP (DW)
Consists of operational current & detailed data	Consists of historical & summarized data
application-oriented. Used for business tasks.	Subject-oriented. Used for Analytics & Mining
OLTP DB are isolated as applications	OLAP integrated per subject area (data mart)
Relatively small, data is archived in MB, and GB.	Large amount of data, stored typically in TB, PB
Supports CRUD (Create, Read, Update, Delete)	Supports only Read
It is volatile	It is non-volatile
Normalized Schema (3NF)	Denormalized Schema

What are the processes that can be done in the data warehouse?

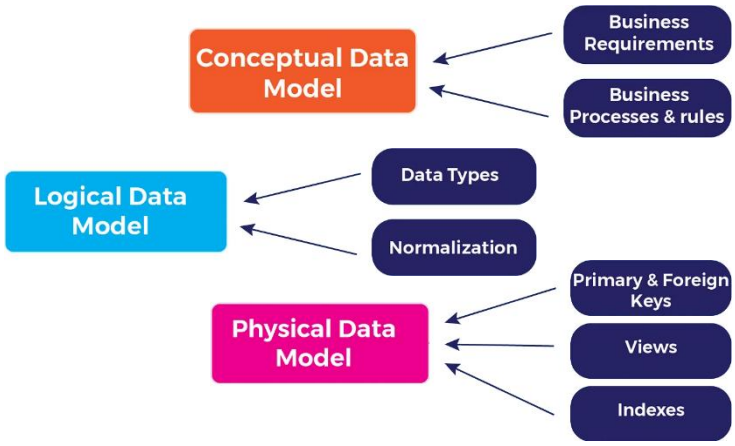
1. Data Extraction:
   extracting data from various data sources, such as databases, applications, and other sources, and transforming it into a format that can be loaded into the data warehouse.
2. Data Transformation:
   cleaning, filtering, merging, and transforming the data extracted from various sources to ensure consistency and accuracy.
3. Data Loading:
   loading the transformed data into the data warehouse. This can be done through various methods such as batch processing, real-time data integration, or incremental loading.
4. Data Aggregation:
   summarizing or consolidating the data into meaningful information that can be used for analysis and reporting.
5. Data Analysis:
   querying and analyzing the data in the data warehouse to gain insights and make informed decisions. This can be done through various tools such as SQL, OLAP, and data mining.
6. Data Visualization:
   presenting the analyzed data in a visual format such as charts, graphs, and dashboards, to help stakeholders understand the information easily.

Data modeling

The process of developing a visual representation of an entire information system or sections to express connections between data points and structures

Types of data modeling

1. Conceptual Data Model
2. Logical Data Model
3. Physical Data Model



Data warehouse modeling

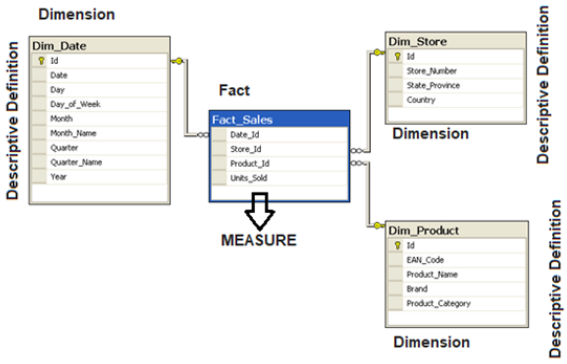
the process of designing the schema or structure of the data warehouse, including the tables, columns, relationships, and constraints that will be used to store and organize the data.

The goal of data modeling in a data warehouse is to create a structure that is optimized for reporting and analysis, and that can support complex queries and aggregations.

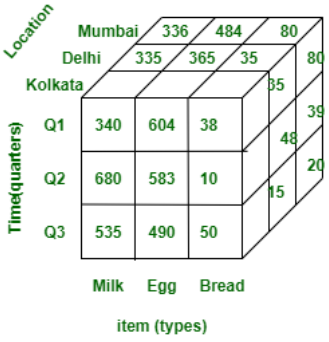
Types of Data warehouse modeling

1- Entity-Relationship (ER) modeling
2- Data Vault modeling
3- Enterprise Multi-Dimensional (EDW) modeling

the most commonly used approach in data warehousing.
It involves organizing data into dimensions and facts.



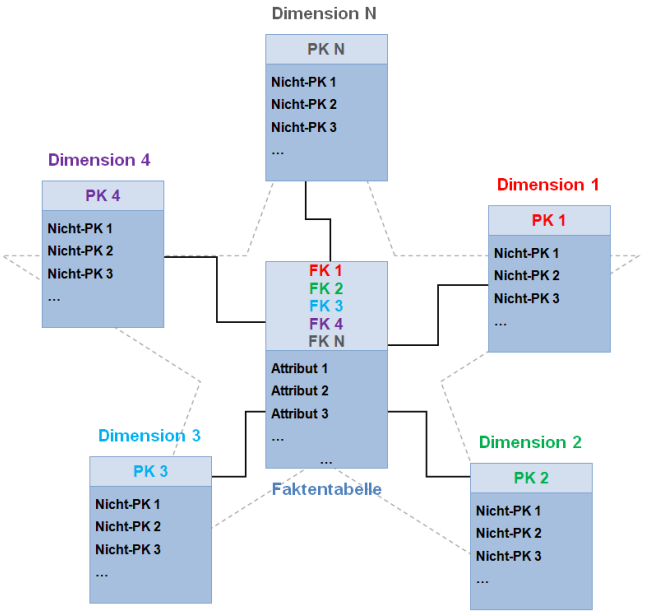
- Dimensions tables (Descriptive Definition)
  represent the various attributes of the data such as time, location, and product.
- Fact table (Numerical Quantities)
  represent the measures or metrics that will be analyzed, such as sales or revenue
- Data cube (data structure)
  represent the multidimensional relationships between measures and dimensions. They provide a fast and efficient way to retrieve and analyze data.



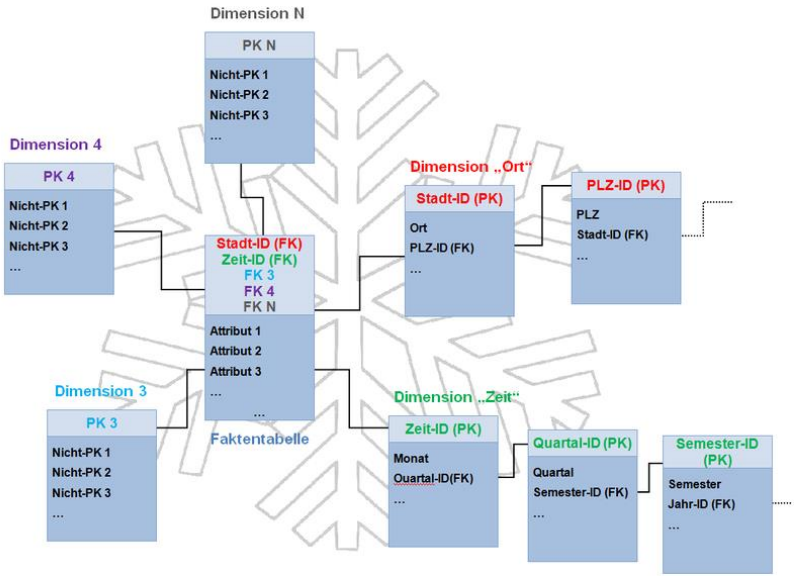
3D data cube
Represented in
2D data table

	Location="Kolkata"			Location="Delhi"			Location="Mumbai"		
	item			item			item		
Time	Milk	Egg	Bread	Milk	Egg	Bread	Milk	Egg	Bread
Q1	340	604	38	335	365	35	336	484	80
Q2	680	583	10	684	490	48	595	594	39
Q3	535	490	50	389	385	15	366	385	20

Multi-Dimensional Modelling Techniques



VS



Star Schema

The most common technique and basic modelling type and is easy to understand. In which Fact table connects with all Dimension tables. Used to develop DWH and Data marts

Snowflake Schema

An extension of the Star Schema where that the dimensional table is normalized into multiple lookup tables, resulting in a more complex structure, in snowflake schema not all dimension tables are directly related to fact table

Star Schema

- more redundant data, difficult to change or maintain.
- less complex, easy to understand.
- All dimension tables are directly connected to the fact table
- fewer foreign keys, query execution is faster and takes lesser time.
- Better for one to one, or one to many relationships
- Both the fact table and dimension tables are denormalized.
- follows a top-down approach.

Snowflake Schema

- less redundant data, easier to change and maintain
- more complex, difficult to understand.
- Not all dimension tables are directly connected to the fact table
- more foreign keys, query execution is slower and takes more time
- Better for complex relationships i.e., many to many relationships
- fact table is denormalized, while dimension tables are normalized.
- follows a bottom-up approach.