

A decorative graphic on the left side of the slide. It consists of a blue parallelogram and a light green parallelogram, both tilted at an angle. The blue shape is in the foreground, and the green shape is partially behind it. They are set against a dark blue background with faint, lighter blue diagonal stripes.

What is Data Warehouse?



Contents

- **Introduction,**
- **Characteristics,**
- **Architecture,**
- **Scheme and modelling,**
- **Operational database systems and data warehouse.**



What is Data Warehouse?

- Data Warehouse is an environment, not a product.
- Data is often scattered across different database, it need DW to get complete information.
- It is aimed at effective integration of operational databases that enables strategic use of data.



Data Warehousing - Introduction

- Data Warehousing *integrates data and information* collected from various sources into one comprehensive database.
(E.g.) *Customer information* from organization's point-of-sale systems, its mailing lists, website and comment cards, etc.
- Data Warehouse is a *centralized storage system* or central repository for storing, analyzing information and interpreting of data in order to facilitate better decision making.
- A data warehouse is a type of data management system that facilitates and supports *business intelligence (BI)* activities, specifically analysis.
- It is primarily designed to facilitate *searches and analyses*
- usually contain large amounts of historical data.



Data Warehouse Usages

- ***Investment & Insurance Companies***– to analyze customer & market trends and allied data patterns.
- ***Retail Chains***– used for marketing and distribution to track items, examine pricing policies and analyze buying trends of customers.
- ***Healthcare***– to generate treatment reports, share data with insurance companies & medical units.
- ***Airline***– operation purpose like crew assignment, route profitability, frequent flyer program promotions, etc.
- ***Banking***– to manage resources available on desk effectively.
- ***Public Sector***– used for intelligence gathering, to maintain & analyze tax records, health policy records, etc.
- ***Telecommunication***– used for product promotions, sales decisions and to make distribution decisions.

Other names for Data Warehouse





Data Warehouse - Definition

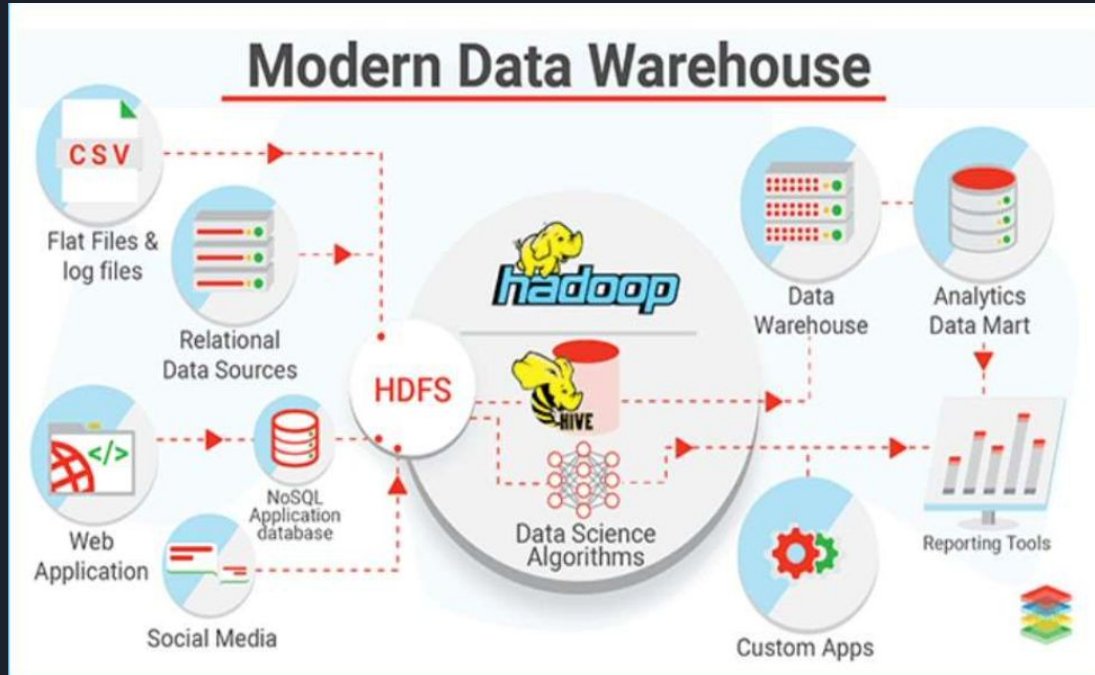
“ A data warehouse is a *single, complete and consistent store* of data obtained from a variety of sources and made available to end users in a way they can understand and use in a business context.”

“ A data warehouse is a *collection of corporate information* derived directly from operational systems and some external data sources.”

“A data warehouse is a *subject-oriented, integrated, time-variant* and non-volatile collection of data in support of management’s decision making process”. - *William H.Inmon*

A data warehouse can be defined as a collection of *organizational data and information* extracted from operational sources and external data sources.

Modern Data Warehouse





How Data Warehouse works?

- Data Warehouse works as a *central repository*, where information arrives from one or more data sources.
- Data flows into *data warehouse from transactional system* and other relational databases.
- Data may be *structured, semi-structured and unstructured* data.
- It is processed, transformed and ingested through *BI tools, SQL* clients and spreadsheets.
- Data warehouse contain *multiple databases* and data is organized into *tables and columns* in each database.
- Data stored in various tables described by *schema & Query tools* use the schema to determine which data tables to access and analyze.
- Data stored in column describe the data such as *integer, data field*, string, etc.
- Data warehousing makes data mining possible.



Characteristics of Data Warehouse

1. Subject-Oriented: It provides *topic wise information* rather than the overall processes of a business. (i.e.) sales, inventory, promotion, etc.

2. Integrated: DW is developed by *integrating data* from varied source into a consistent format. It is stored in data warehouse in a consistent manner in terms of *naming, format & coding*, which facilitates data analysis.

3. Non-Volatile: Data once entered into a data warehouse must *remain unchanged* & all data is read only.

4. Time Variant: Data stored in a data warehouse is documented with an *element of time*, either explicitly or implicitly. It is exhibited in primary key with element of time like *day, week*, etc.



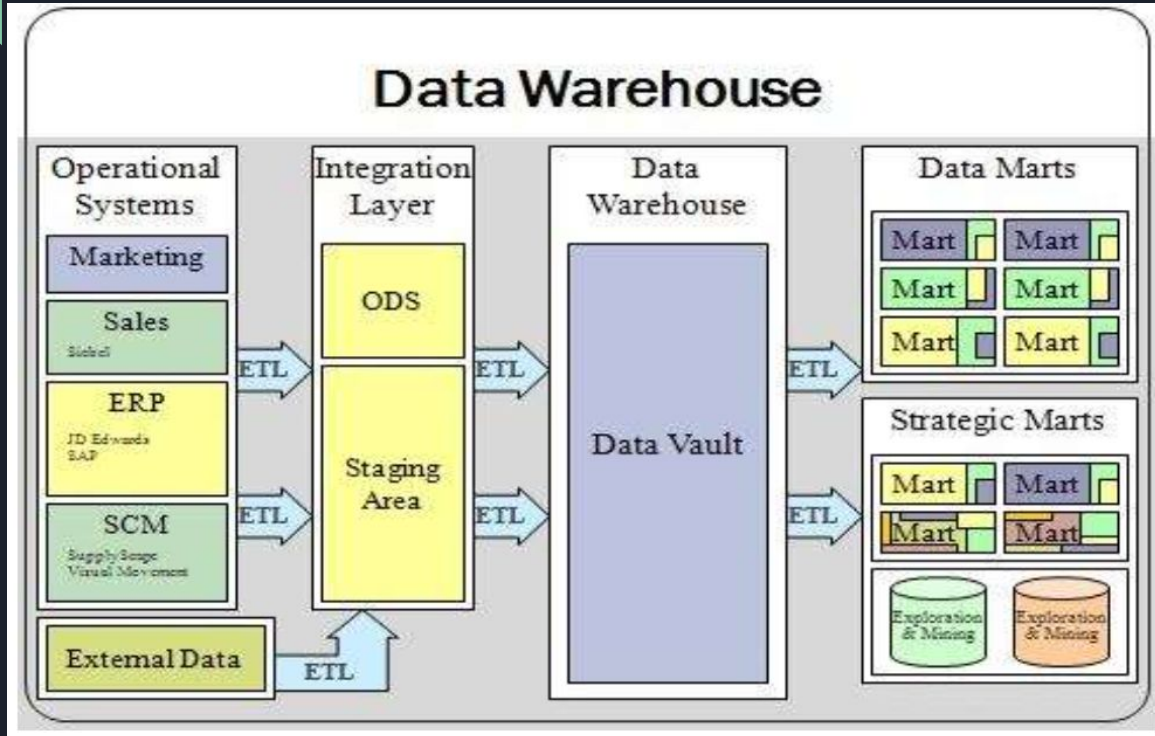
Types of Data Warehouse

1. Enterprise Data Warehouse (EDW): It serves as a *key or central database* that facilitates decision support services throughout the enterprise. It provides access to *cross-organizational information*, offers a unified approach to data representation.

2. Operational Data Store (ODS): It is preferred for *routine activities* like storing employee records.

3. Data Mart: It is a *subset of a data warehouse* built to maintain a particular department, region or business unit. *Every department* of a business has a central repository or data mart to store data.

Data Warehouse example



Data Warehouse Vs Database

Characteristics	Data Warehouse	Transactional Database
Suitable workloads	Analytics, reporting, big data	Transaction processing
Data source	Data collected and normalized from many sources	Data captured as-is from a single source, such as a transactional system
Data capture	Bulk write operations typically on a predetermined batch schedule	Optimized for continuous write operations as new data is available to maximize transaction throughput.

Data Warehouse Vs Database

Characteristics	Data Warehouse	Transactional Database
Data normalization	Denormalized schemas, such as the Star schema or Snowflake schema	Highly normalized, static schemas
Data storage	Optimized for simplicity of access and high-speed query performance using columnar storage	Optimized for high throughput write operations to a single row-oriented physical block
Data access	Optimized to minimize I/O and maximize data throughput	High volumes of small read operations

Data Warehouse Vs Data lake

Characteristics	Data Warehouse	Data lake
Data	Relational data from transactional systems, operational databases, and line of business applications	All data, including structured, semi-structured, and unstructured
Schema	Often designed prior to the data warehouse implementation but also can be written at the time of analysis (schema-on-write or schema-on-read)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using local storage	Query results getting faster using low-cost storage

Data Warehouse Vs Data lake

Characteristics	Data Warehouse	Data lake
Data quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (i.e. raw data)
Users	Business analysts, data scientists, and data developers	Business analysts (using curated data), data scientists, data developers, data engineers, and data architects
Analytics	Batch reporting, BI, and visualizations	Machine learning, exploratory analytics, data discovery, streaming, operational analytics, big data, and profiling



Data Warehouse – Components

The four components of data warehouse are–

- 1. Load Manager (Front Component):** It performs with all the operations associated with the *extraction and load of data* into warehouse.
- 2. Warehouse Manager:** It performs operations associated with the management of data in the warehouse. It performs operations like analysis of data to ensure consistency, creation of *indexes & views*, *transformation*, merging of source data.
- 3. Query Manager (Backend Component):** It performs all operation related to management of *user queries*.
- 4. End-User access tools:** It is categorized into 5 groups like *data reporting*, *query tools*, application development tools, *EIS tools*, *OLAP tools* and data mining tools.



Data Warehouse – Tools

Data Warehouse applications can be categorized as:

- Query and reporting tools.
- Application Development tools.
- Data Mining tools.
- OLAP tools.

Some *popular data warehouse tools* are -

- Xplenty.
- Amazon Redshift.
- Teradata.
- Oracle 12c.
- Informatica.
- IBM Infosphere.
- Cloudera.
- Panoply.



Benefits of Data Warehouse

There are *several benefits of data warehouse* for end users like:

- Improved data consistency, data quality and accuracy.
- Better business decisions.
- Easier access to enterprise data for end-users.
- Better documentation of data.
- Historical data analysis.
- Reduced computer costs and higher productivity.
- Enabling end-users to ask ad-hoc queries or reports without deterring the performance of operational systems.
- Collection or consolidation of related data from various sources into a place.



Advantages of Data Warehouse

- It allows business users to *quickly access critical data* from some sources.
- It provides consistent information on various *cross-functional activities*.
- It helps to integrate *many sources of data* to reduce stress on production system.
- It helps to reduce total turnaround time for *analysis & reporting*.
- It stores a *large amount of historical data* and helps users to analyse different time periods.
- *Restructuring & Integration* make it easier for user to use for reporting & analysis.

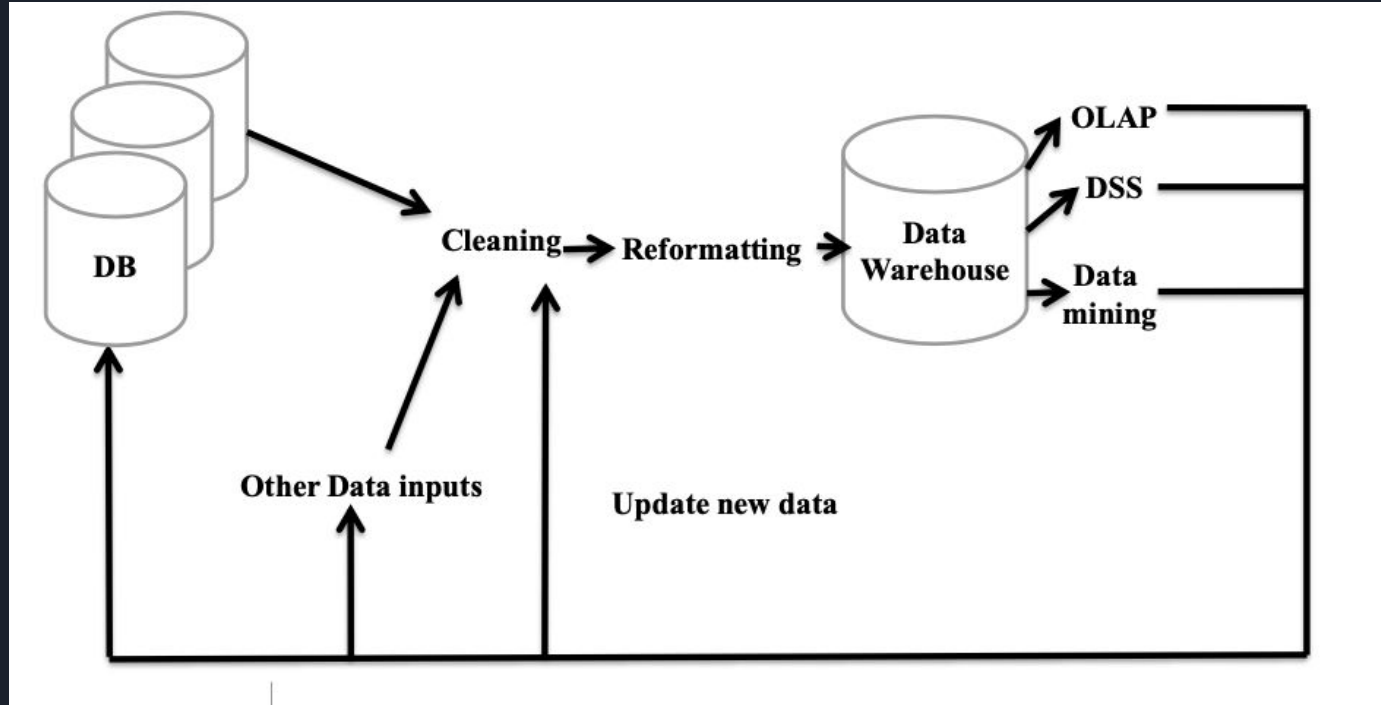


Disadvantages of Data Warehouse

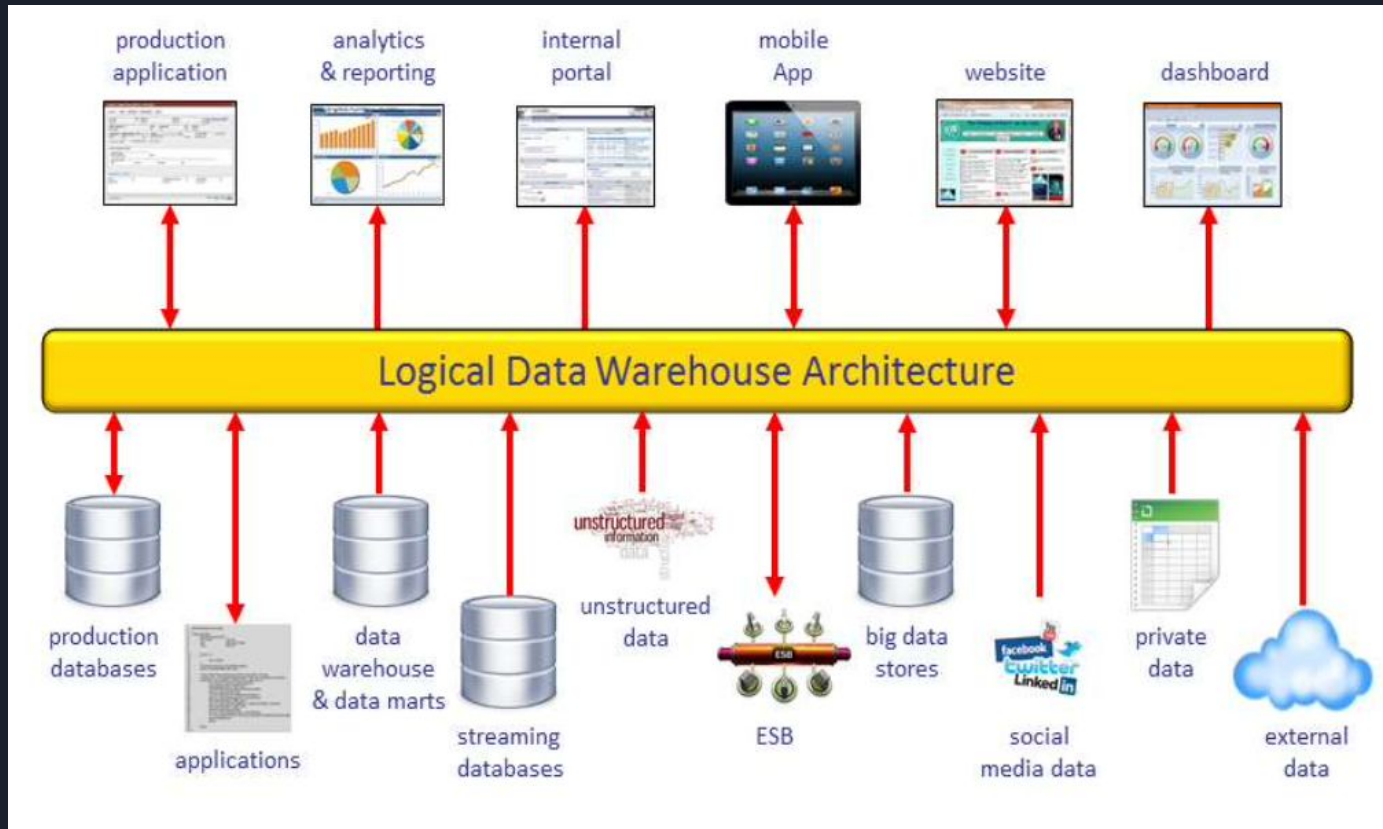
- Data Warehouse can be *outdated relatively quickly*.
- Difficult to *make changes* in data types and ranges, data source schema, indexes and queries.
- Data warehouse seem easy, but actually, it is *too complex* for average users.
- Sometime warehouse users will develop *different business rules*.
- Organization need to spend *lots of their resources* for training and implementation purpose.
- This is not an ideal option for *unstructured data*.

Structure of Data Warehouse (Architecture)

1. Basic Structure (Single Tier):

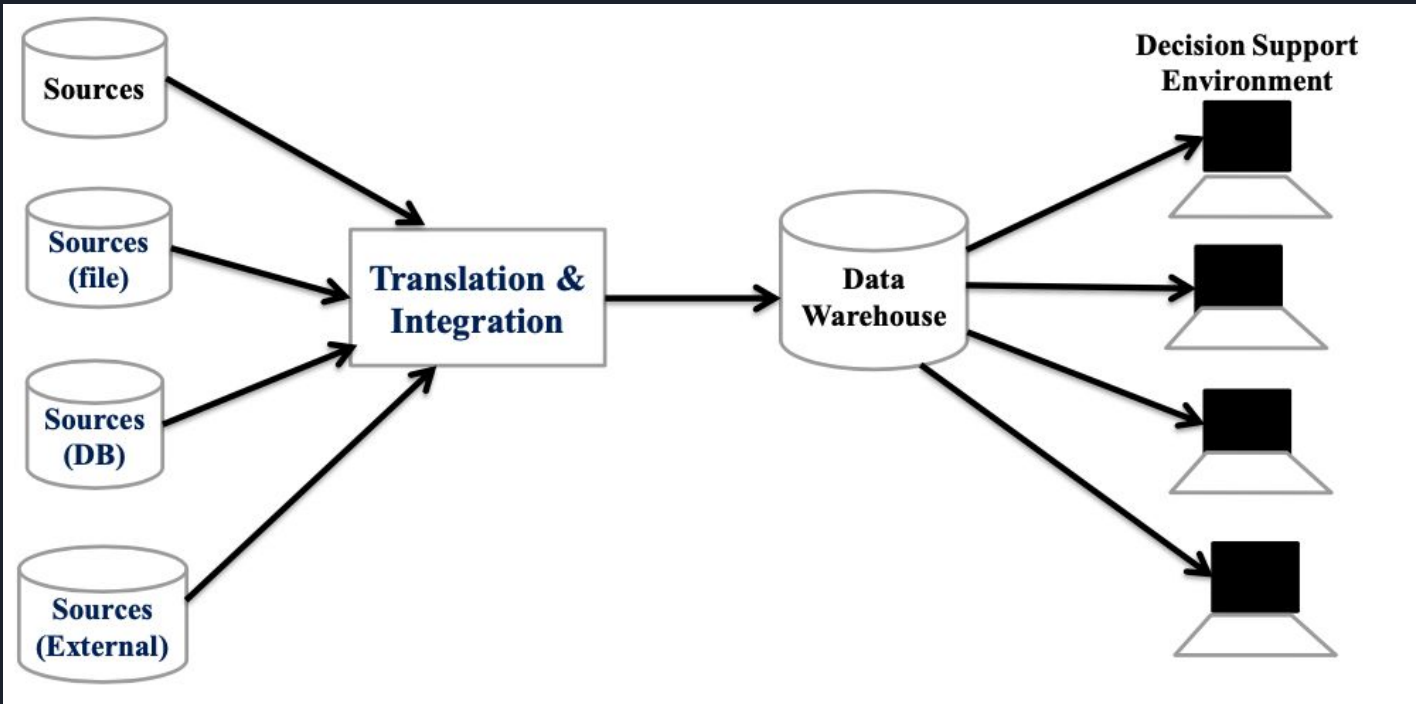


Sample Data Warehouse Architecture



Two Tier Architecture

2. Generic Two Level (Two Tier):





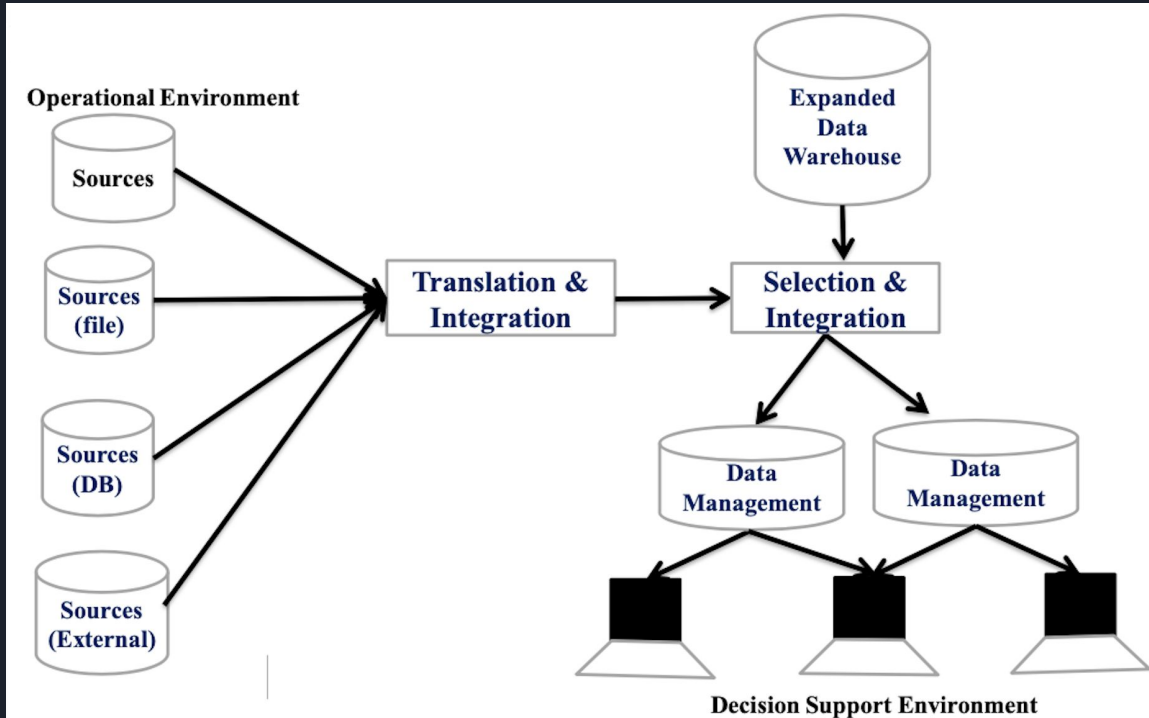
Two Tier (Level) Architecture:

Two-layer architecture is one of the Data Warehouse layers which separates physically available ***sources and data warehouse***. This architecture is ***not expandable*** and also not supporting a large number of end-users. It also has connectivity problems because of network limitations.

A two-tier architecture includes a ***staging area*** for all data sources, before the data warehouse layer. By adding a staging area between the sources and the storage repository, you ensure all data loaded into the warehouse is ***cleansed*** and in the appropriate format.

Three Tier Architecture

3. Generic Three Level (Three Tier):





Three Tier (Level) Architecture:

This is the most widely used Architecture of Data Warehouse. It consists of the Top, Middle and Bottom Tier.

a) **Bottom Tier:**

- It is the *database* of the Data warehouse servers.
- It is usually a *relational database* system.
- Data is *cleansed, transformed*, and loaded into this layer using back- end tools.

It includes–

- Data Extraction**– get data from multiple, heterogeneous, and external sources.
- Data cleaning** - detect errors in the data and rectify them when possible.
- Data transformation** - convert data from legacy or host format to warehouse format.
- Load** - sort, summarize, consolidate, compute views, check integrity, and build indices and partitions.
- Refresh** - propagate the updates from the data sources to the warehouse.



Three Tier (Level) Architecture:

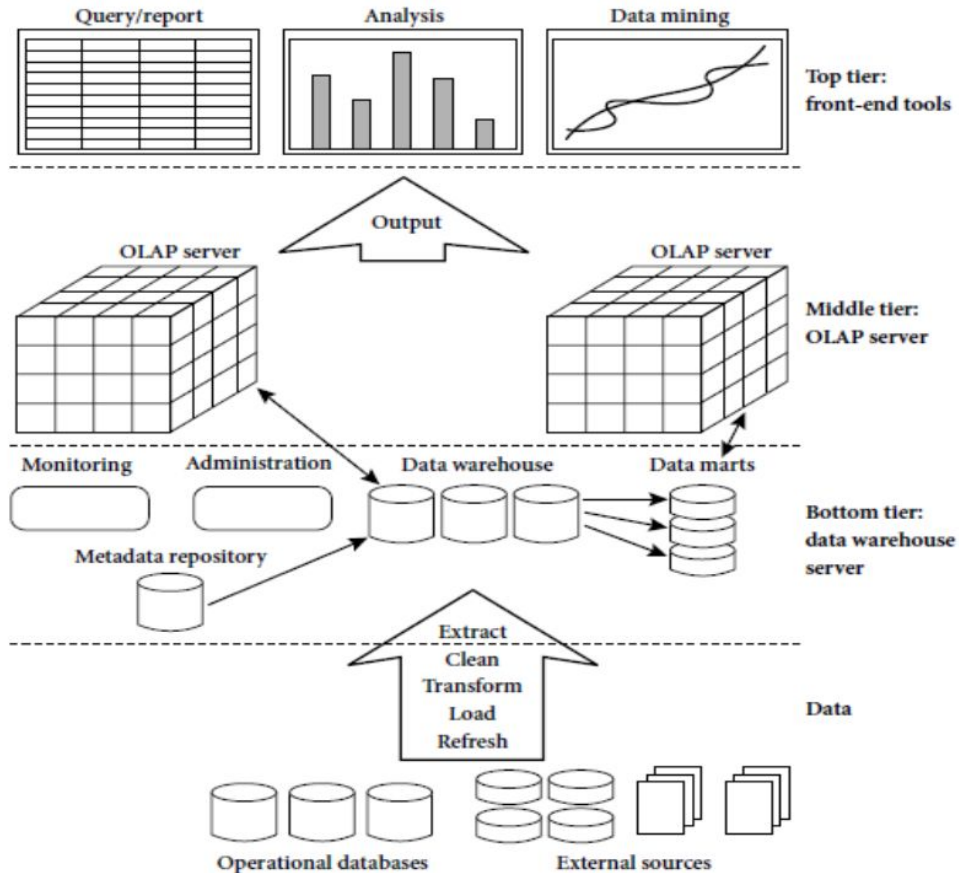
b) Middle Tier:

- It is the *application layer* giving an *abstracted view* of the database.
- The middle tier in Data warehouse is an OLAP server which is implemented using either *ROLAP or MOLAP* model.
- This layer also acts as a mediator between the end-user and the database.
- It arranges the data to make it more *suitable for analysis*.

c) Top-Tier:

- It represent the *front-end client layer*.
- It is where the user *accesses and interacts* with the data.
- It could be *Query tools, reporting tools*, managed query tools, *Analysis tools* and Data mining tools.

Three Tier (Level) Architecture:





Data Warehouse Models

1. Enterprise Warehouse Models:

“An enterprise warehouse models collects *all of the information* about subjects spanning the entire organization”.

- It provides *corporate-wide data integration*, usually one or more operational systems or external information providers and is *cross-functional* scope.
- It consists of *summarized data* and its range in size from gigabytes, terabytes or beyond.
- It is implemented on *traditional mainframes*, computer super servers, parallel architecture platforms.
- It takes years to *design and build*.



Data Warehouse Models

2. Data Mart:

“A data mart contains a *subset of corporate-wide data* that is of value to a specific group of users”.

(E.g.) *Marketing data mart* contains details about customer, item and sales.

- It is implemented on *low– cost* departmental servers (i.e.) Unix / Linux.
- It involve *complex integration* in long run.
- It is categorized into two source of data–

i) *Independent data marts*– data captured from *one or more operational systems* or external information or geographic area.

ii) *Dependent data marts*– it is sourced directly from *enterprise data warehouses*.



Data Warehouse Models

3. *Virtual Warehouse Models:*

“A Virtual warehouse models is a *set of views* over operational databases”.

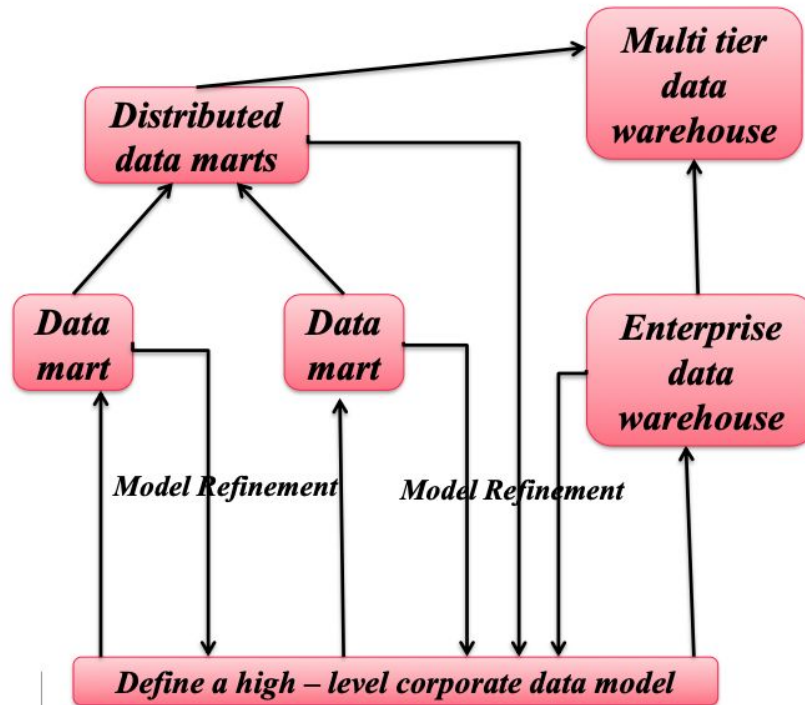
- It is easy to build but requires *excess capacity* on operational database servers.

a) Top-down approach– it is a systematic solution and *minimizes integration* problems.

- It is expensive, *takes long time* to develop, lacks flexibility due to difficulty in achieving *consistency*.

b) Bottom-down approach– it is used to *design, development* and deployment of independent data mart provides flexibility, *low cost* and rapid return of investment.

Example approach for data warehouse development





Meta Data Models

Metadata: “Metadata are *data about data*. It represents *warehouse objects* when used in data warehouse. It is created for the data names and definitions of given data warehouse”.

- It is created and captured for *any extracted data*, source of extracted data and it is added by *cleaning or integration* process.
- *Metadata repository* are placed in the bottom of the data warehouse architecture.
- It is used as directory to *help decision support system* analyst locate the contents of data warehouse.
- It serve as a guide to *algorithms* used for summarization.
- Metadata should be stored and managed persistently.

A metadata repository consists of–

a) Data warehouse structure– it includes data warehouse schema, view, *dimensions, hierarchies* and derived data definitions as well as data mart locations and contents.



Meta Data Models

b) Operational metadata– it include *data lineage* (i.e.) history of migrated data and sequence of transformations), currency of data and monitoring information (error reports and audit trails).

c) Algorithms used for summarization– includes measure and dimension definition *algorithms, data on partitions*, subject areas, aggregation, summarization and predefined queries and reports.

d) Mapping from operational environment to data warehouse– includes databases and their *contents, descriptions*, data partitions, data extraction, cleaning and security.

e) Data related to system performance– includes indices and profiles that *improve data access* and retrieval performance, scheduling of update and replication cycles.

f) Business metadata– it includes business terms and definitions, *data ownership information* and charging policies.



Data Warehouse Modelling

Data warehouse Modeling:

- Data warehouse and OLAP tools are based on a *multidimensional data model*.
- This model views in the form of a *data cube*.

Schemas for Multidimensional Model: “A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by *dimensions and facts*”.

Dimensions– it is perspectives or *entities* with respect to which an organization wants to keep records. (E.g.) Sales data warehouse has dimensions like time, item, branch and location.

- It is specified by *users or experts* and adjust data.



Data Warehouse Modelling

Facts— it is numeric measures and analyze the relationships between dimensions. (E.g.) Sales data warehouse include sales amount in terms of dollars sold, units sold.

- The data cube is a ***metaphor*** for multidimensional data storage.
- The actual storage of such data may differ from its ***logical representation***.
- The data at different degrees of summarization is referred as ***Cuboid***.
- The cuboid that holds the lowest level of summarization is called the ***base cuboid***.
- The cuboid that holds the highest level of summarization is called the ***apex cuboid***.



Schemas

Stars, Snowflakes and Fact Constellations: Schemas for Multidimensional Data models

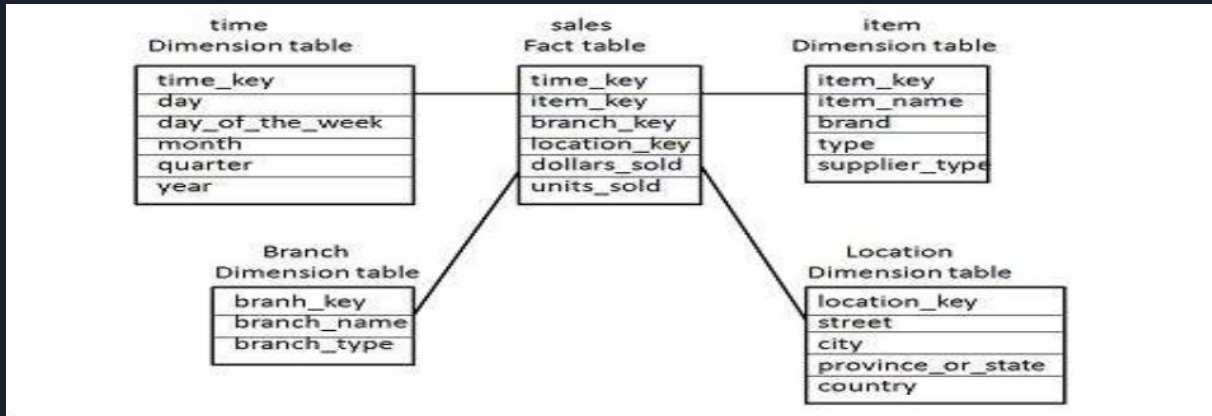
- ER data model is commonly used in design of relational databases, where *database schema* consists of a set of entities and the relationships between them.
- The most popular data model for a data warehouse is *multidimensional data* that consists of Star schema, Snowflake schema and Fact constellation schema.

a) Star Schema:

- It establishes the relationships between the *fact table* and any one of the dimension tables.
- It has single fact table connected to dimension tables *like a star*.
- It has one fact table and is *associated with numerous dimensions*
- table and depicts a star.

Schemas

(E.g.) *Star Schema diagram for sales data of a company with respect to four dimensions: time, item, branch & location.*



- Each dimension represented with only *one dimension table*.
- Dimension table contains *set of attributes*.
- Fact table at centre & it contains key to each of *four dimensions*.
- Fact table contains attributes like, *dollars sold* & units sold.



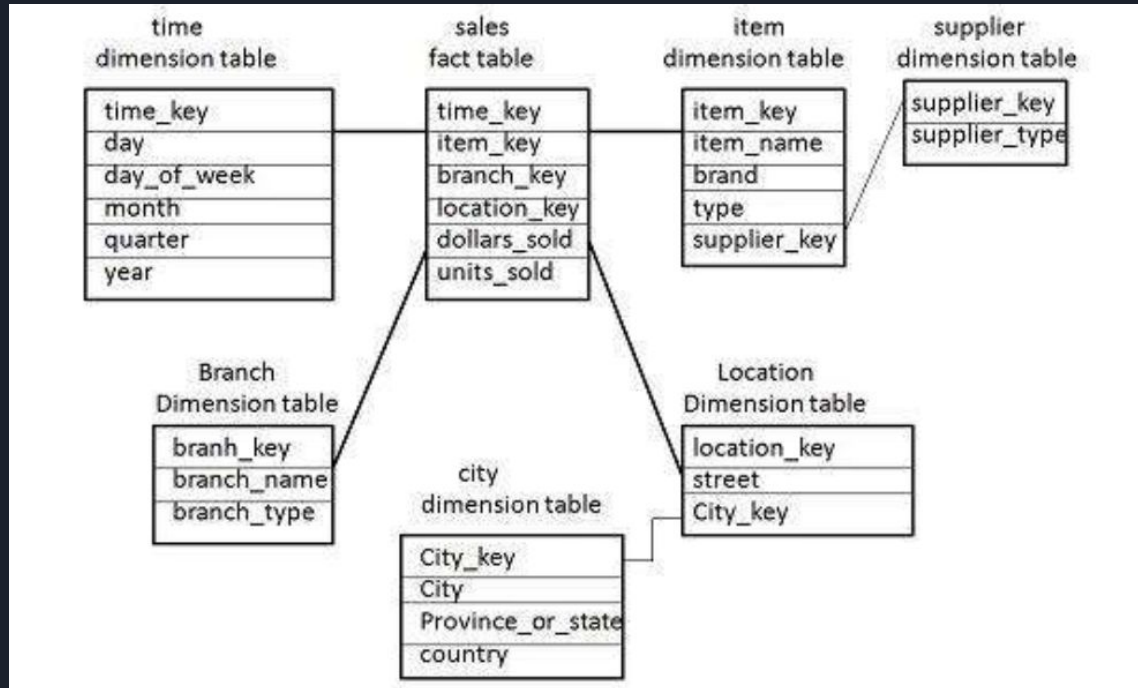
Schemas

b) Snowflake Schema:

- It is the *extension of star schema*.
- Large dimension tables are normalized into *multiple tables*.
- The normalization split up the data into *additional tables*.
- There is relationship between the dimensions tables & it has to do joins to fetch the data. In diagram, item dimension table is normalized and split into two dimension tables namely, item and supplier table.
- Item dimension table contains attributes item_key, item_name, type, brand and supplier_key.
- Supplier dimension table contains the attributes supplier_key and supplier_table.

Schemas

(E.g.) Snowflake Schema diagram for sales data of a company:





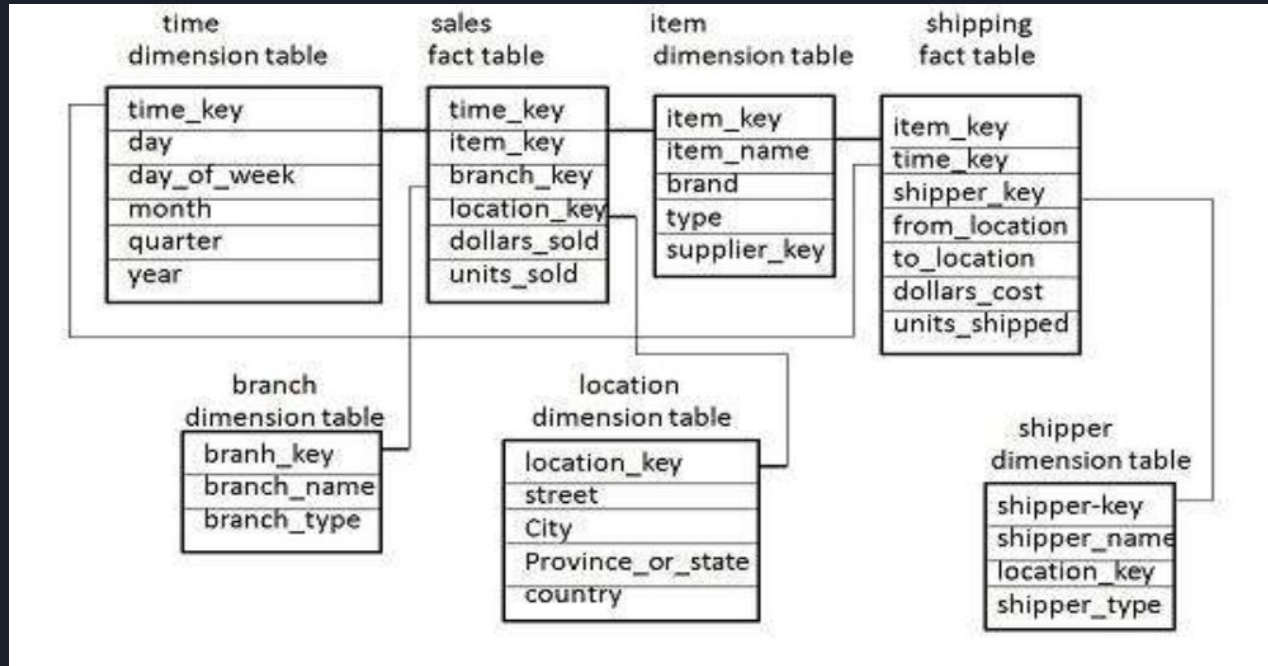
Schemas

c) Fact constellation:

- Fact constellation is a measure of *online analytical processing*, which is collection multiple fact tables sharing dimension tables, viewed as a collection of stars.
- It consists of *multiple fact tables* & this schema is also known as galaxy schema.
- In following diagram, it consists of *two fact tables*, sales and shipping.
- Sale fact table is same as star schema.
- Shipping fact table has five dimensions namely, item_key, time_key, shipper_key, location.
- It contains two measures namely dollars sold and units sold.
- It is possible for dimension table to share between fact tables.

Schemas

(E.g.) Fact Constellation Schema diagram for sales data of a company:




Difference between Star, Snowflake & Fact Constellation Schema

Characteristics	Star Schema	Snowflake Schema	Fact Schema
Elements	Single Fact Table connected to multiple dimension tables with no sub-dimension tables	Single Fact Table connects to multiple dimension tables that connects to multiple sub-dimension tables	Multiple Fact Tables connects to multiple dimension tables that connects to multiple sub-dimension tables
Normalization	Denormalized	Normalized	Normalized
Number of Dimensions	Multiple dimension tables map to a single Fact Table	Multiple dimension tables map to multiple dimension tables	Multiple dimension tables map to multiple Fact Tables


Difference between Star, Snowflake & Fact Constellation Schema

Characteristics	Star Schema	Snowflake Schema	Fact Schema
Data Redundancy	High	Low	Low
Performance	Fewer foreign keys resulting in increased performance	Decreased performance compared to Star Schema from higher number of foreign keys	Decreased performance compared to Star and Snowflake. Used for complex data aggregation.
Complexity	Simple, designed to be easy to understand	More complicated compared to Star Schema – can be more challenging to understand	Most complicated to understand. Reserved for highly complex data structures



Difference between Star, Snowflake & Fact Constellation Schema

Characteristics	Star Schema	Snowflake Schema	Fact Schema
Storage Usage	Higher disk space due to data redundancy	Lower disk space due to limited data redundancy	Low disk space usage compared to the level of sophistication due to the limited data redundancy
Design Limitations	One Fact Table only, no sub-dimensions	One Fact Table only, multiple sub-dimensions are permitted	Multiple Fact Tables permitted, only first level dimensions are permitted



Operational Database systems and Data warehouse

Operational Database systems:

- Operational system assist a company or organization in its *day-to- day business*.
- It's applications and data are highly structure and provide immediate focus on *business functions* with the help of OLTP.
- It required to support a *large number of transaction* on a daily basis.
- Operational data stores *small, focusing the database* on specific business area and eliminating database overhead in areas such as indexes.

Data Warehouse system:

- This system is organized *around the trends* or patterns in those events set by operational systems.
- Data warehouse *focus on business needs* and requirements.
- It develop ideas for *changing the business rules* to make these events more effective



Operational Database systems and Data warehouse

The multiple purposes are:

- It minimises the impact of reporting and *complex query processing* on operational systems.
- It preserves *operational data* for reuse after that data has been purged from the operational systems.
- It manages the *data based on time*, allowing the user to look back and see how the company looked in the past versus the present.
- It provides a data store that can be modified to conform to the way the *users view* the data.
- It unifies the data within a common business definition, offering *one version of reality*.

Difference between Operational Database systems and Data warehouse

Difference	Operational System	Data Warehouse system
Size & Content	Small	Large
Performance	Speed in nature	Slow, as the request
Content focus	Small work areas	Cross functional
Tools	Typical structure and deals with less no. of tools	Various tools and supports the types of data



Thank you for your attention!