Unit-1

Introduction of Data Warehousing

❖ What is Data Warehousing?

- Data warehouse is process of constructing & using a data warehousing.
- A data warehouse integrates by data from multiple heterogeneous sources that support analytical report, structure & adhoc queries & decision making.
- "A data warehouse is a collection of corporate information & data derived from operational systems & external data source."
- Data is populated into data warehouse through the process of extraction, transformation & loading.
- **A data warehouse is a large collection of business data used to help an organization make decisions.**
- A Data warehouse is typically used to connect and analyze business data from heterogeneous sources.
- It is a process of transforming data into information and making it available to users in a timely manner to make a difference.
- Suppose a business executive wants to analyze previous feedback on any data such as a product, a supplier, or any consumer data, then the executive will have no data available to analyze because the previous data has been updated due to transactions.
- Data warehouse also provides Online Analytical Process (OLAP).

❖ Understanding of Data Warehousing

- A data warehouse is a database, which is kept separate from the organization's operational database.
- There is no frequent updating done in a data warehouse.
- It possesses consolidated historical data, which helps the organization to analyze its business.
- A data warehouse helps executives to organize, understand, and use their data to take strategic decisions.
- Data warehouse systems help in the integration of diversity of application systems.
- A data warehouse system helps in consolidated historical data analysis.

❖History of Data Warehousing

- Data warehouse was first developed by **Bill Inmon in 1990.**
- According to Inmon a data warehouse is a subject oriented, integrated, time variant & non-volatile collection of data.
- However, the real concept was given by Inmon Bill. He was considered as a father of data warehouse. He had written about a variety of topics for building, usage, and maintenance of the warehouse & the Corporate Information Factory.
- Data warehouse takes data from many different sources to a single location & translated into a format the data warehouse can process & store.
- A data warehouse help execute to organize, understand & use their data to take strategy decision.

❖ Data Warehousing Today/How Data WareHouse Work ?

- A Data Warehouse works as a central repository where information arrives from one or more data sources. Data flows into a data warehouse from the transactional system and other relational databases.

  Data may be:

  - **Structured**
  - **Semi-structured**
  - **Unstructured data**

- The data is processed, transformed, and ingested so that users can access the processed data in the Data Warehouse through Business Intelligence tools, SQL clients, and spreadsheets.
- A data warehouse merges information coming from different sources into one comprehensive database.
- By merging all of this information in one place, an organization can analyze its customers more holistically. This helps to ensure that it has considered all the information available.
- Data warehousing makes data mining possible. Data mining is looking for patterns in the data that may lead to higher sales and profits.
- Data warehouse technology is fast & new tool to take care of our future needs.
- Data warehousing combine information collected for multiple sources into one comprehensive database.
- Data warehouse subject oriented because it provides information around a subject rather than organization ongoing operations.
- Data warehouse is essential database & is kept separate from an

organization operational database.

What Is a Data Warehouse Used For?

Here, are most common sectors where Data warehouse is used:

- **Airline:**

  In the Airline system, it is used for operation purpose like crew assignment, analyses of route profitability, frequent flyer program promotions, etc.

- **Banking:**

  It is widely used in the banking sector to manage the resources available on desk effectively. Few banks also used for the market research, performance analysis of the product and operations.

- **Healthcare:**

  Healthcare sector also used Data warehouse to strategize and predict outcomes, generate patient's treatment reports, share data with tie-in insurance companies, medical aid services, etc.

- **Public sector:**

  In the public sector, data warehouse is used for intelligence gathering. It helps government agencies to maintain and analyze tax records, health policy records, for every individual.

- **Investment and Insurance sector:**

  In this sector, the warehouses are primarily used to analyze data patterns, customer trends, and to track market movements.

- **Retain chain:**

  In retail chains, Data warehouse is widely used for distribution and marketing. It also helps to track items, customer buying pattern, promotions and also used for determining pricing policy.

- **Telecommunication:**

A data warehouse is used in this sector for product promotions, sales decisions and to make distribution decisions.

- **Hospitality Industry:**

This Industry utilizes warehouse services to design as well as estimate their advertising and promotion campaigns where they want to target clients based on their feedback and travel patterns.

➢ Different form of a data warehouse

1) Information processing data warehouse.
2) Analytical processing data warehouse.
3) Data mining data warehouse.

1. Information processing data warehouse

- This specially allows processing of historical data which if store in it.
- There are many processing operations which can be performed like query, generating tables, chart or graph & basic operation under statistic analysis.

2. Analytical processing data warehouse

- This warehouse can be used for extensive analytical processing & this analytics performed on the data is stored in the warehouse.
- Which perform OLAP operational & few others like drill down & drill up with enhance the result of the analysis.

3. Data mining data warehouse

- This warehouse is dedicated to the data mining the discovery of information by uncovering, hidden pattern, prediction technique & analytical model construction.
- Data warehouse (DWH) today system follow updates approach rather than the traditional discuss.

❖ Data warehouse today many tools & utilities

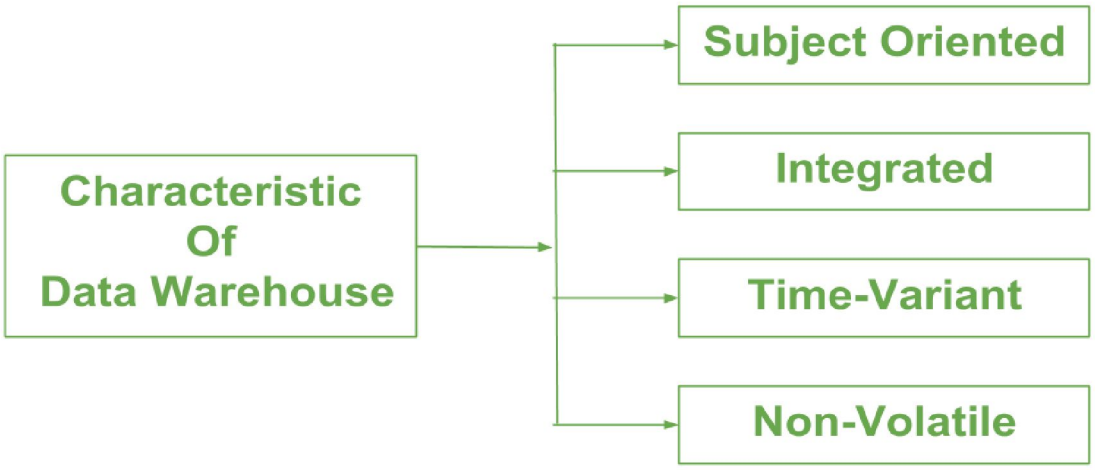
- Data Extraction.
- Data Cleaning.
- Data Transformation.
- Data Loading.

❖ Today data warehouse latest tool

- Amazon red shift.
- Tera data.
- Panoply.
- Oracle 12C.
- Informatica.
- IBM Infopshere.
- Paraccel.
- Cloudera.
- Marklogic.
- SAP.

❖ Feature/Characteristics of warehousing

- Data warehouse can be controlled when the user has a shared way of explaining the trends that are introduced as specific subject. Below are major Feature/Characteristics of data warehouse:



- **Subject Oriented** – A data warehouse is subject oriented because it provides information around a subject rather than the organization's ongoing operations. These subjects can be product, customers, suppliers, sales, revenue, etc. A data warehouse does not focus on the ongoing operations; rather it focuses on modeling and analysis of data for decision making.

  A data warehouse is always a subject oriented as it delivers information about a theme instead of organization's current operations. It can be achieved on specific theme. That means the data warehousing process is proposed to handle with a specific theme which is more defined. These themes can be sales, distributions, marketing etc.

- **Integrated –** A data warehouse is constructed by integrating data from heterogeneous sources such as relational databases, flat files, etc. This integration enhances the effective analysis of data.

  It is somewhere same as subject orientation which is made in a reliable format. Integration means founding a shared entity to scale the all similar data from the different databases. The data also required to be resided into various data warehouse in shared and generally granted manner.

- **Time Variant** − The data collected in a data warehouse is identified with a particular time period. The data in a data warehouse provides information from the historical point of view.

  In this data is maintained via different intervals of time such as weekly, monthly, or annually etc. It founds various time limit which are structured between the large datasets and are held in online transaction process (OLTP). The time limits for data warehouse is wide-ranged than that of operational systems.

- **Non-volatile** − Non-volatile means the previous data is not erased when new data is added to it. A data warehouse is kept separate from the operational database and therefore frequent changes in operational database are not reflected in the data warehouse.

  In this, data is read-only and refreshed at particular intervals. This is beneficial in analyzing historical data and in comprehension the functionality. It does not need transaction process, recapture and concurrency control mechanism.

  Functionalities such as delete, update, and insert that are done in an operational application are lost in data warehouse environment. Two types of data operations done in the data warehouse are:

  - Data Loading
  - Data Access

❖Future trends in data warehouse

- Data warehouse has Online Analytical Processing (OLAP) technology are now considered mature.
- The data warehouse has been the business insight work of enterprise computing.
- It means emerging technology in data warehouse & analytics in cloud.
- Data warehousing & big data analysis are emerging area in cloud today.
- Data warehouse one or more has a part of an enterprise data warehouse the hub is a subset of the data called the data core.
- Data warehouse to support two stage:
    - Refined data.
    - Trusted data.
- Data warehouse architecture cannot stay alone we must think purpose & position of the data warehouse in data management architecture.
- The major demand in data warehouse market for high speed data mining at lower hardware and implementation cost.
- In a future of data warehouse modernization we will need to consider cloud data warehousing with automation as well as architecture modernization.
- Data warehousing services lead the information management category in increase rate jumping for 24% to 34% in 2014.
- As we have seen that the size of the open database has grown approximately double its magnitude in the last few years, it shows the significant value that it contains.
- As the size of the databases grows, the estimates of what constitutes a very large database continue to grow.
- The hardware and software that are available today do not allow keeping a large amount of data online. For example, a Telco call record requires 10TB of data to be kept online, which is just a size of one month's record. If it requires keeping records of sales, marketing customer, employees, etc., then the size will be more than 100 TB.

- The record contains textual information and some multimedia data. Multimedia data cannot be easily manipulated as text data. Searching the multimedia data is not an easy task, whereas textual information can be retrieved by the relational software available today.
- Apart from size planning, it is complex to build and run data warehouse systems that are ever increasing in size. As the number of users increases, the size of the data warehouse also increases. These users will also require accessing the system.
- With the growth of the Internet, there is a requirement of users to access data online.
- Visual analytics makes the work a little bit more pleasant, but **machine learning** (predictive analytics) is game changing.
- Data warehouses have hundreds or thousands of data points per customer. It is not possible to manually look at each factor. It is certainly not possible to look at every combination of two or three factors.
- More complex impact factors – such as how monthly income is important for customers under 40, for older customers the gender is the key factor – are completely impossible to figure out manually.
- Data warehouse staying power because the concept of central data collects by dozens or hundreds of database, applications & system other source system.
- To the most efficient way of companies to get an enterprise wide view of their customers, supply chain, sales& operations.
- In today world of instant access by many different areas user & customers data is no longer nicely away in big warehouse.
- The trend is two words always on accessible & very open storage that is fast & friendly for customers yet complex & deep in of the most important data.

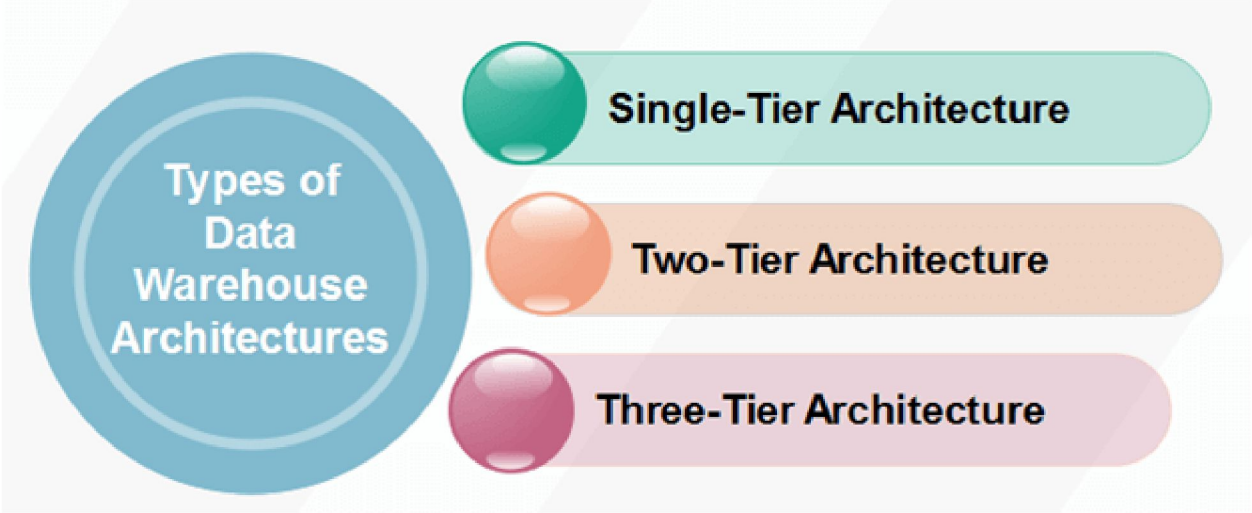
❖Data warehouse Architecture

- Data warehouse architecture refers to the design of an organization's data collection and storage framework. Because data needs to be sorted, cleaned, and properly organized to be useful, data warehouse architecture focuses on finding the most efficient method of taking information from a raw set and placing it into an easily digestible structure that provides valuable Business Intelligent.
- Data warehouse architecture is a method of defining the overall architecture of data communication processing and presentation that exist for end-clients computing within the enterprise.
- Data warehouses and their architectures very depending upon the elements of an organization's situation.
- A data warehouse is the source of business truth development by combined data from multiple separate sources.
- It supports analytical, reporting & both structure & adhoc queries.
- **There are three types of architecture:**

  1) Single Tier Architecture.

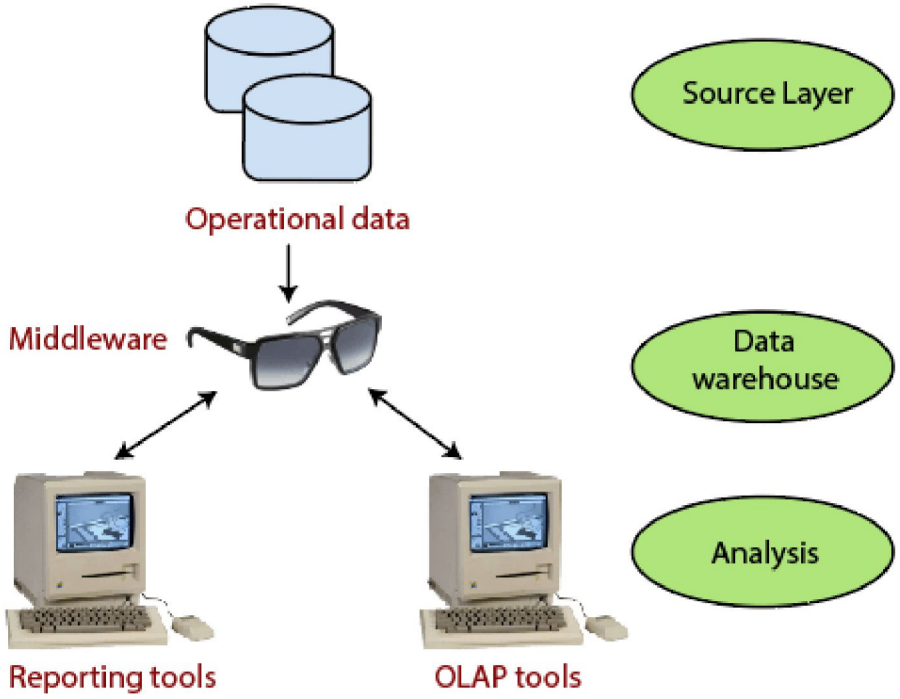
  2) Two Tier Architecture.

  3) Three Tier Architecture.



1) Single Tier Architecture

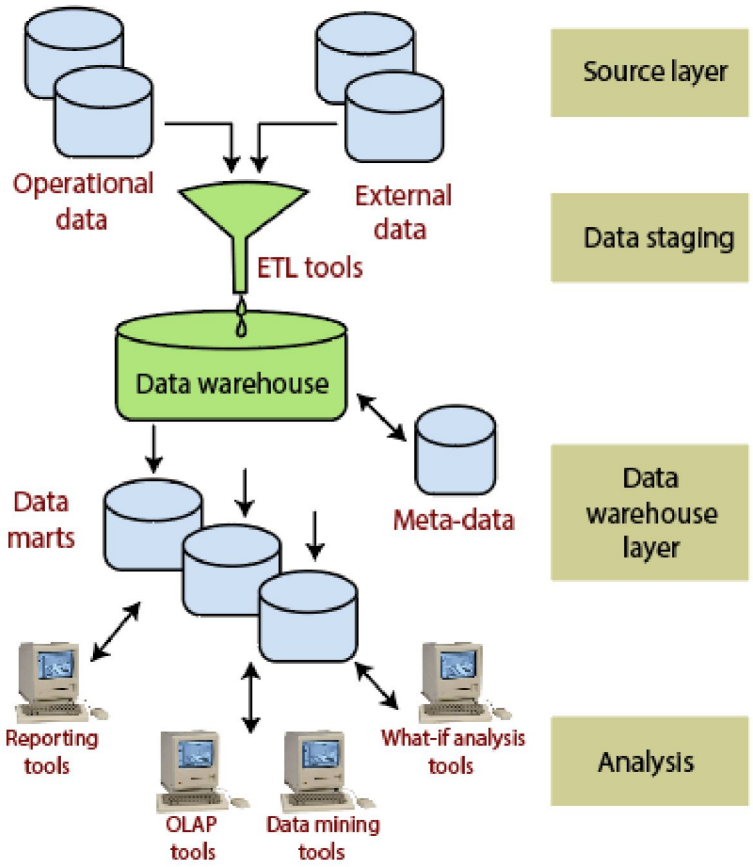
- The object of a single layer is to minimize the amount of data store.
- This goal is to remove data redundancy.
- This architecture is not frequently used.
- The figure shows the only layer physically available is the source layer. In this method, data warehouses are virtual. This means that the data warehouse is implemented as a multidimensional view of operational data created by specific middleware, or an intermediate processing layer.



Single-Tier Data Warehouse Architecture

2) Two Tier Architecture

- Two tier architecture separate physically available sources & data warehouse.
- This architecture is not expandable & also not supporting a large number of end users.
- It also has connectivity problems because of network limitations.



Two-Tier Data Warehouse Architecture

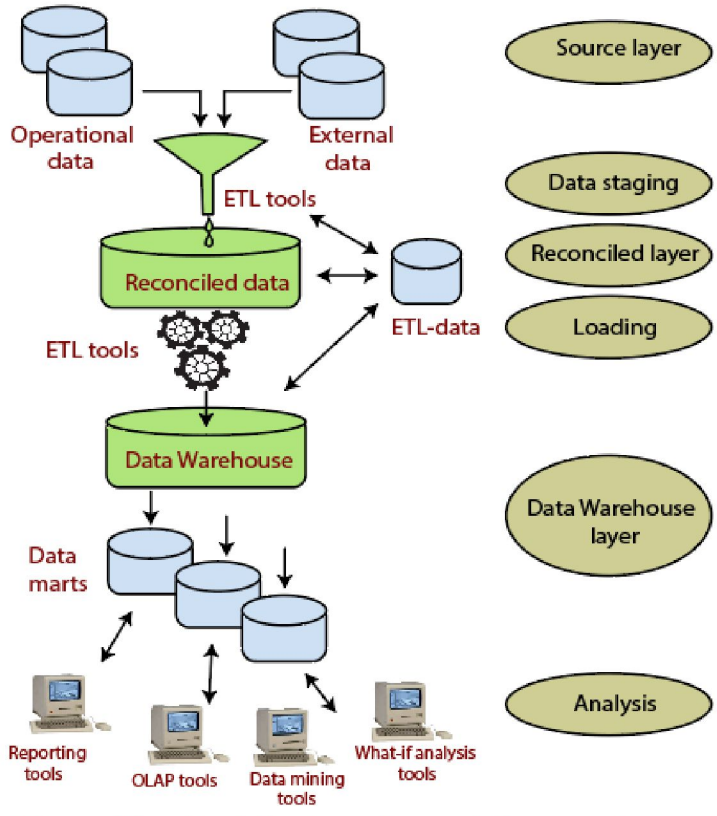
- Although it is typically called two-layer architecture to highlight a separation between physically available sources and data warehouses, in fact, consists of four subsequent data flow stages:
- **Source layer:** A data warehouse system uses a heterogeneous source of data. That data is stored initially to corporate relational databases or legacy databases, or it may come from an information system outside the corporate walls.
- **Data Staging:** The data stored to the source should be extracted, cleansed to remove inconsistencies and fill gaps, and integrated to merge heterogeneous sources into one standard schema. The so-named Extraction, Transformation, and Loading Tools (ETL) can combine heterogeneous schemata, extract, transform, cleanse, validate, filter, and load source data into a data warehouse.
- **Data Warehouse layer:** Information is saved to one logically centralized individual repository: a data warehouse. The data warehouses can be directly accessed, but it can also be used as a source for creating data marts, which partially replicate data

warehouse contents and are designed for specific enterprise departments. Meta-data repositories store information on sources, access procedures, data staging, users, data mart schema, and so on.

- **Analysis:** In this layer, integrated data is efficiently, and flexible accessed to issue reports, dynamically analyze information, and simulate hypothetical business scenarios. It should feature aggregate information navigators, complex query optimizers, and customer-friendly GUIs.

3) Three Tier Architecture

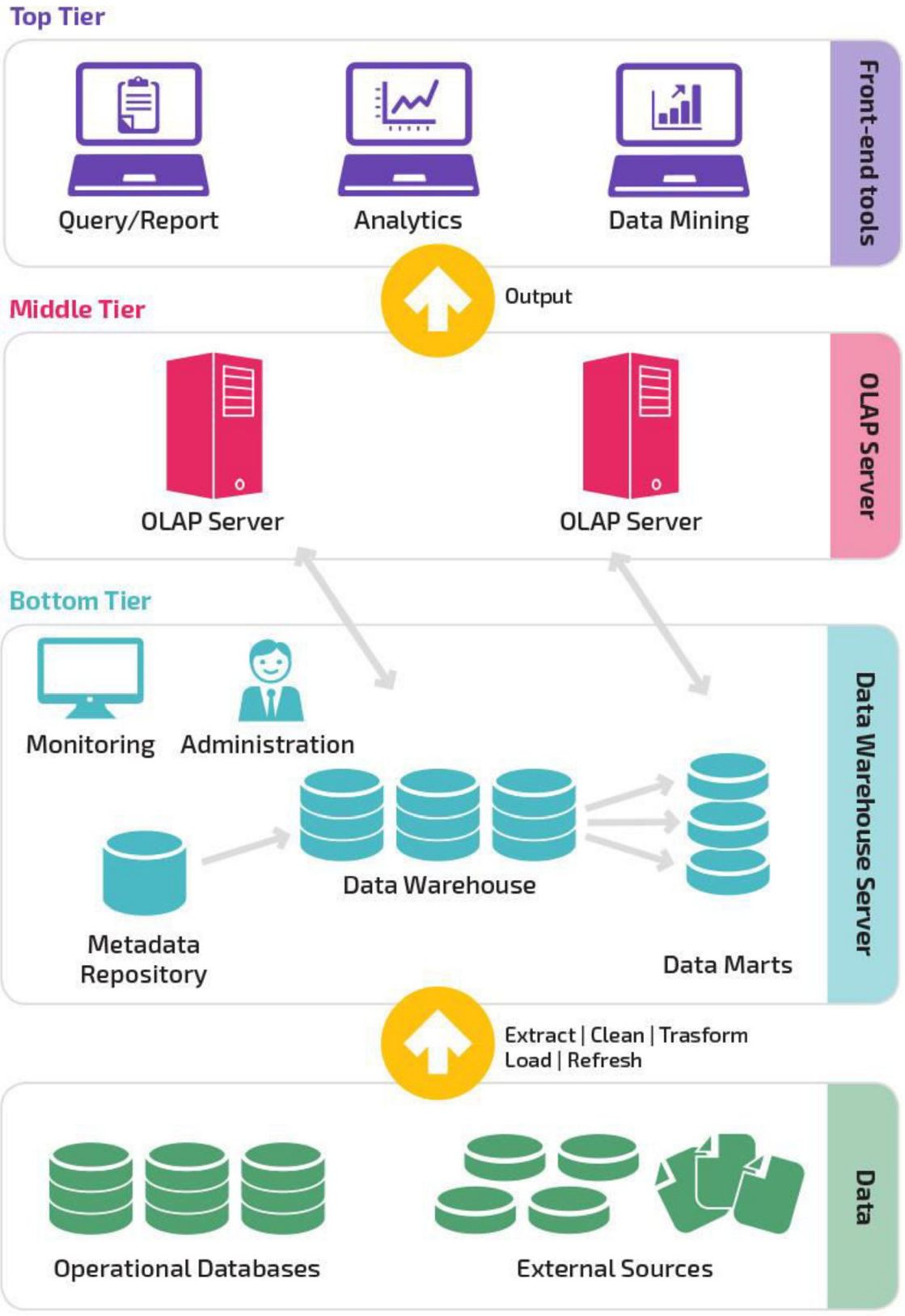
- This is the most widely used architecture. This architecture is especially useful for the extensive, enterprise-wide systems. A disadvantage of this structure is the extra file storage space used through the extra redundant reconciled layer. It also makes the analytical tools a little further away from being real-time.
- It consists of top, middle & bottom tier.



Three-Tier Architecture for a data warehouse system

Data Warehouses usually have a three-level (tier) architecture that includes:

1. **Bottom Tier (Data Warehouse Server)**
2. **Middle Tier (OLAP Server)**
3. **Top Tier (Front end Tools).**



1. Bottom Tier (Data Warehouse Server)

- The bottom tier of the architecture is the data warehouse database server.
- We use the back-end tools & utilities to feed data in the bottom tiers.
- This back-end tools & utilities perform the extract, clean, load & refresh function (DWH) tools.
- Data warehouse server fetch relevant information based on data mining & request.
- A **bottom-tier** that consists of the **Data Warehouse server**, which is almost always an RDBMS. It may include several specialized data marts and a metadata repository.
- Data from operational databases and external sources (such as user profile data provided by external consultants) are extracted using application program interfaces called a gateway. A gateway is provided by the underlying DBMS and allows customer programs to generate SQL code to be executed at a server.
- **Examples** of gateways contain **ODBC** (Open Database Connection) and **OLE-DB** (Open-Linking and Embedding for Databases), by **Microsoft**, and **JDBC** (Java Database Connection).

2. Middle Tier (OLAP Server)

- A **middle-tier** which consists of an **OLAP server** for fast querying of the data warehouse. The middle tier in data warehouse in an OLAP server which is implemented using either ROLAP & MOLAP model.
- For a user this application tier presents and abstract view of database.
- This layer also acts as a mediator between the end user and database.

The OLAP server is implemented using either

(1) A **Relational OLAP (ROLAP) model**, i.e., an extended relational DBMS that maps functions on multidimensional data to standard relational operations.

(2) A **Multidimensional OLAP (MOLAP) model**, i.e., a particular purpose server that directly implements multidimensional information and operations.

3. TOP Tier (Front End Tools)

- The top tier is a front end client layer.
- Top tier is the tool & API that connect & get data out from the data warehouse.
- It could be query tools, reporting tools, manage query tool, analysis tool & data mining tools.
- A **top-tier** that contains **front-end tools** for displaying results provided by OLAP, as well as additional tools for data mining of the OLAP-generated data.

- Establishing which type of database your organization needs and how you plan to interact with it is searching for insights.
- It is also important to evaluate who is going to be examining data and what sources they need when considering your data warehouse architecture.
- Although the data warehouse debate is not always applicable for smaller organizations, those with more teams, departments, and specific needs may benefit from the latter.
- Finally, Architecture depending on the size of your organization and different types of warehouse architectures may be more practical. Understanding which is best depends on the currency of your data, the size of your sets, and your organization's demands.

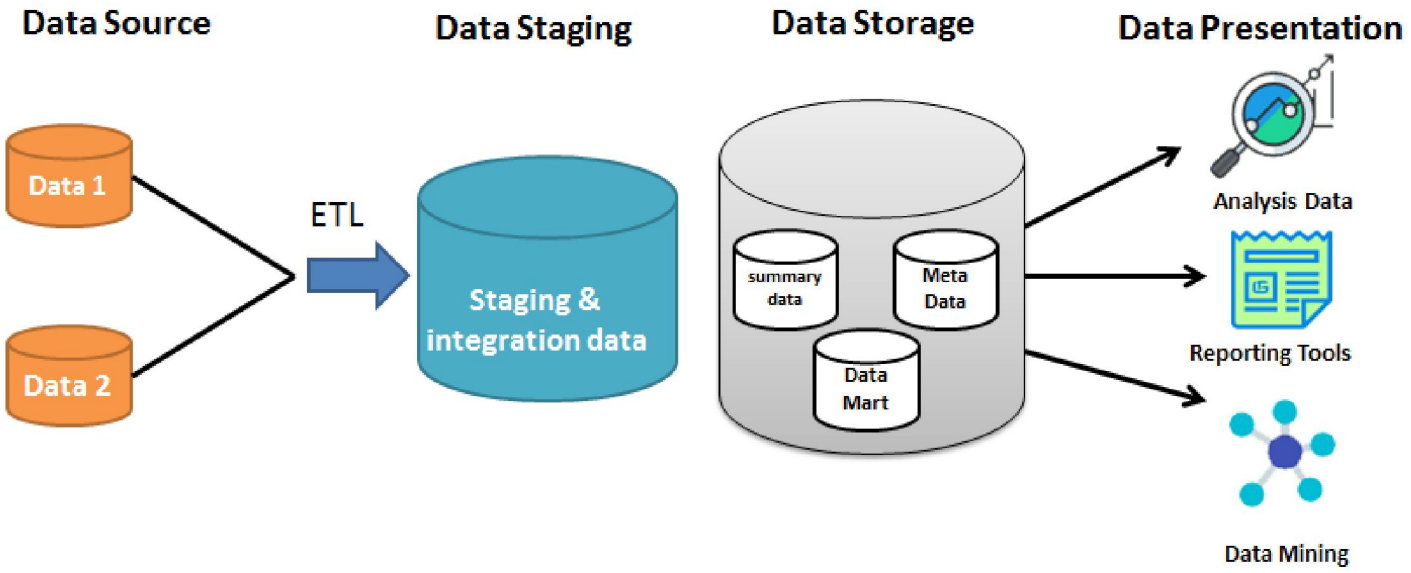
➢ Data warehouse architecture includes the following layers:

1) Data Source Layer.

2) Data Storage Layer.

3) Data Presentation Layer.

4) Data Staging Layer.



1. Data Source Layer

- The Data Source Layer is the layer where the data from the source is encountered and subsequently sent to the other layers for desired operations.
- The data can be of any type.
- The Source Data can be a database, a Spreadsheet or any other kinds of a text file.
- The Source Data can be of any format. We cannot expect to get data with the same format considering the sources are vastly different.
- In Real Life, Some examples of Source Data can be
- Log Files of each specific application or job or entry of employers in a company.
- Survey Data, Stock Exchange Data, etc.
- Web Browser Data and many more.

2. Data Staging Layer

The following steps take place in Data Staging Layer.

Step #1: Data Extraction

The Data received by the Source Layer is feed into the Staging Layer where the first process that takes place with the acquired data is extraction.

Step #2: Landing Database

- The extracted data is temporarily stored in a landing database.
- It retrieves the data once the data is extracted.

Step #3: Staging Area

- The Data in Landing Database is taken and several quality checks and staging operations are performed in the staging area.
- The Structure and Schema are also identified and adjustments are made to data that are unordered thus trying to bring about a commonality among the data that has been acquired.
- Having a place or set up for the data just before transformation and changes is an added advantage that makes the Staging process very important.
- It makes data processing easier.

Step #4: ETL

- It is an Extraction, Transformation, and Load.
- ETL Tools are used for integration and processing of data where logic is applied to rather raw but somewhat ordered data.
- This data is extracted as per the analytical nature that is required and transformed to data that is deemed fit to be stored in the Data Warehouse.

- After Transformation, the data or rather information is finally loaded into the data warehouse.
- Some examples of ETL tools are Informatica, SSIS, etc.

- The data staging layer resides between data source and data warehouse.
- In this layer data is extracted from different internal and external data source.
- The data extraction layer will utilize multiple technologies & tools to extract the required data.
- The extract has been loaded it will be subjected to high level data quality data checks.
- **The staging layer contain the following components:**

  1) Landing Database & Staging Area.

  2) Data Integration Tool.

- A data warehouse relational database that is designed for query & analysis rather than for transaction processes.
- It usually contains historical data derived from transaction data but can include data from other sources.

3. Data Storage Layer

- The processed data is stored in the Data Warehouse.
- This Data is cleansed, transformed, and prepared with a definite structure and thus provides opportunities for employers to use data as required by the Business.
- Depending upon the approach of the Architecture, the data will be stored in Data Warehouse as well as Data Marts. Data Marts will be discussed in the later stages.
- Some also include an Operational Data Store.

- Data storage layer is where data was cleaned in storage area as a single central process.

4. Data Presentation Layer

- This Layer where the users get to interact with the data stored in the data warehouse.
- Queries and several tools will be employed to get different types of information based on the data.
- This layer of data warehouse architecture provides users with the ability to query the data for product or services insight, analyzed &information to business scenarios & developed automated or adhoc reports.
- You may OLAP or reporting tool with a user friendly Graphical User Interface (GUI) to help users build their queries perform analysis or designed their reports.
- The information reaches the user through the graphical representation of data.
- Reporting Tools are used to get Business Data and Business logic is also applied to gather several kinds of information.
- Meta Data Information and System operations and performance are also maintained and viewed in this layer.

❖Data Flow Architecture

- A data warehouse system has two mains architecture.
    1) **The Data Flow Architecture.**
    2) **The System Architecture.**
- The data flow architecture how the data store are arranged within a data warehouse & how the data flow from the source system to the users through this data source.
- The system architecture is about the physical configuration of the server, network software & clients.
- The data flow architecture is a configuration of data stores within a data warehouse system.
- This includes how the data flow is controlled logged & monitored as well as the mechanism to ensure the quality of the data in data stores.
- The data flow architecture is different from data architecture.
- Data architecture is about how the data is arranged in each data store & how a data is designed to reflect the business process.
- Data flow in the data warehouse describes which objects are needed at design time& which objects are needed at runtime to transfer data from a source to destination.
- The individual requirements of your company process are supported by number of ways to design the data flow.
- You can use any data source that transfer the data to destination or access the source data directly apply simple & complex methods & define data correspond to the requirements of your layer architecture.
- Data flow architecture transforms input data by a series of computational components into output data.
- Data flow architecture reduces development time & can move easily between design & implementation.
- Data flow architecture is a computer architecture that directly contrasts the traditional architecture or control flow architecture.

- Data flow architecture the data can be input graph topology with a cycle.
- Data flow architecture doesn't have a program counter or execution of instructions is determined based on availability.
- There are benefits and drawbacks to each type of data.
- **For Example:**-The data that you can access in the data warehouse is more complex & can represent a greater number & several of relationship but it can take longer time to collect & access that live data.
- Data flow architecture that are determined in nature enable program to manage complex task such as processor load balancing synchronization & access to common resources.
- Several of process enables reports to access application data in order to produce report output & live reporting data is accessed directly from the application using the API.