Overview on Data Warehouse

- The term "Data Warehouse" was first coined by Bill Inmon in 1990. According to Inmon, a data warehouse is a subject oriented, integrated, time-variant, and non-volatile collection of data. This data helps analysts to take informed decisions in an organization.
- A data warehouses provides us generalized and consolidated data in multidimensional view.
- Along with generalized and consolidated view of data, a data warehouses also provides us Online Analytical Processing OLAP tools.
    - These tools help us in interactive and effective analysis of data in a multidimensional space.
    - This analysis results in data generalization and data mining.
        - Data mining functions such as association, clustering, classification, prediction can be integrated with OLAP operations to enhance the interactive mining of knowledge at multiple level of abstraction.
- 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.

Data Warehouse Features

The key features of a data warehouse are discussed below:

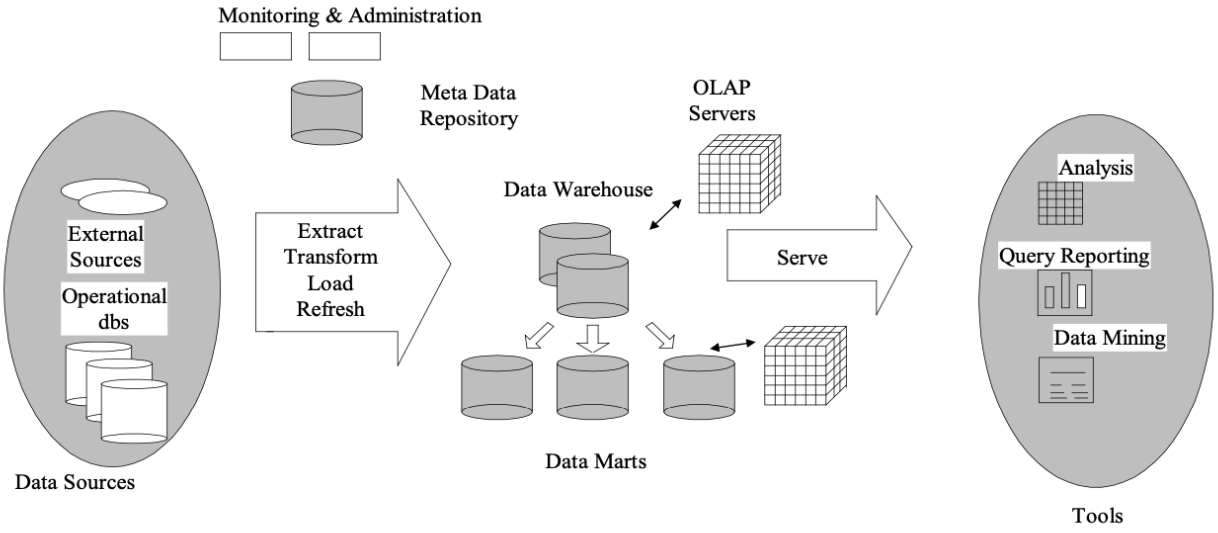
1. **Subject Oriented** - A data warehouse is subject oriented because it provides information around a subject rather than the organization's ongoing operations. These subjects can be

product, customers, suppliers, sales, revenue, etc. A data warehouse does not focus on the ongoing operations, rather it focuses on modelling and analysis of data for decision making.

2. **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.

3. **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.

4. **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 is not reflected in the data warehouse.



Types of Data Warehouse

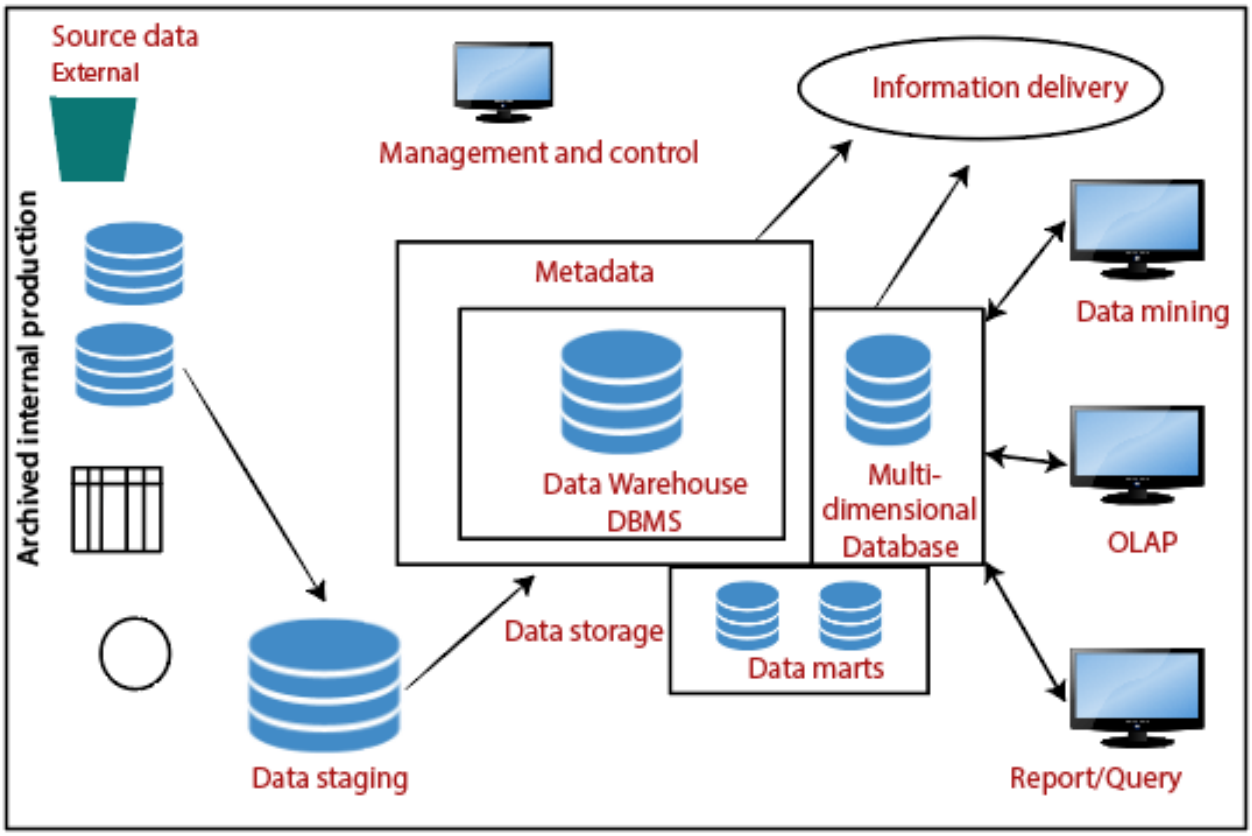
Information processing, analytical processing, and data mining are the three types of data warehouse applications that are discussed below:

1. **Information Processing** - A data warehouse allows to process the data stored in it. The data can be processed by means of querying, basic statistical analysis, reporting using crosstabs, tables, charts, or graphs.

2. **Analytical Processing** - A data warehouse supports analytical processing of the information stored in it. The data can be analyzed by means of basic

OLAP operations, including slice- and-dice, drill down, drill up, and pivoting.

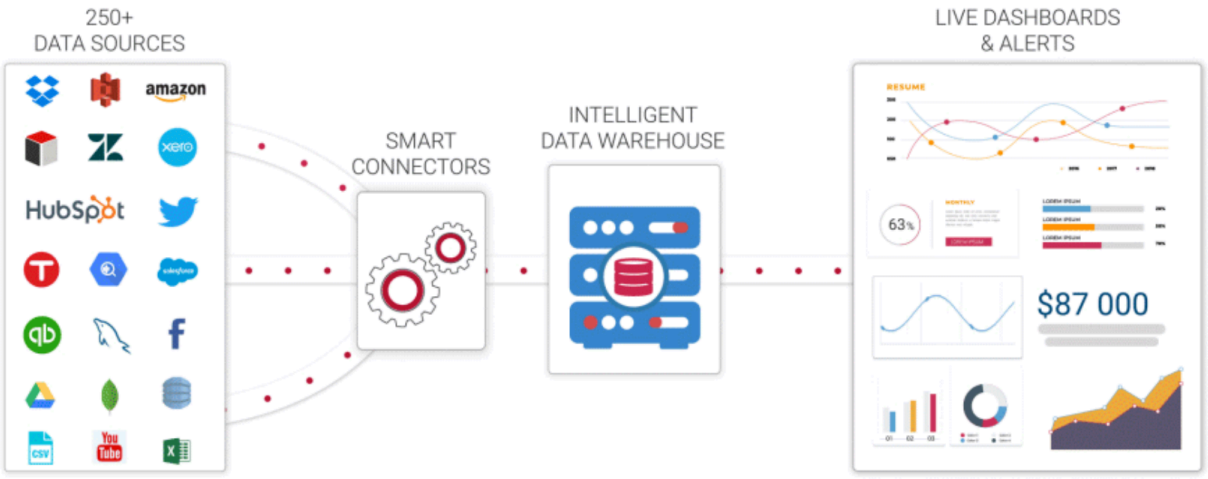
3. **Data Mining** - Data mining supports knowledge discovery by finding hidden patterns and associations, constructing analytical models, performing classification and prediction. These mining results can be presented using the visualization tools.



Components or Building Blocks of Data Warehouse

- A data warehouse (DW) is a central repository where data is stored in query-able forms.
- From a technical standpoint, a data warehouse is a relational database optimized for reading, aggregating, and querying large volumes of data. Traditionally, DWs only contained structured data, or data that can be arranged in tables.
- However, modern DWs can combine both structured and unstructured data where unstructured refers to a wide variety of forms (such as images, pdf files, audio formats, etc.) that are harder to categorize and process.

- Without DWs, data scientists would have to pull data straight from the production database and may wind up reporting different results to the same question or cause delays and even outages.



- Surprisingly, DW isn't a regular database. How so?
- First of all, they differ in terms of data structure. A regular database normalizes data excluding any data redundancies and separating related data into tables. This takes up a lot of computing resources, as a single query combines data from many tables. Contrarily, a DW uses simple queries with few tables to improve performance and analytics.
- Second, aimed at day-to-day transactions, standard transactional databases don't usually store historic data, while for warehouses, it's their main purpose, as they collect data from multiple periods. DW simplifies a data analyst's job, allowing for manipulating all data from a single interface and deriving analytics, visualizations, and statistics.
- Typically, a data warehouse doesn't support as many concurrent users as a database, being designed for a small group of analysts and business users.

Data Warehousing - Terminologies

Data Cube

- A data cube helps us represent data in multiple dimensions. It is defined by dimensions and facts.

- The dimensions are the entities with respect to which an enterprise preserves the records.

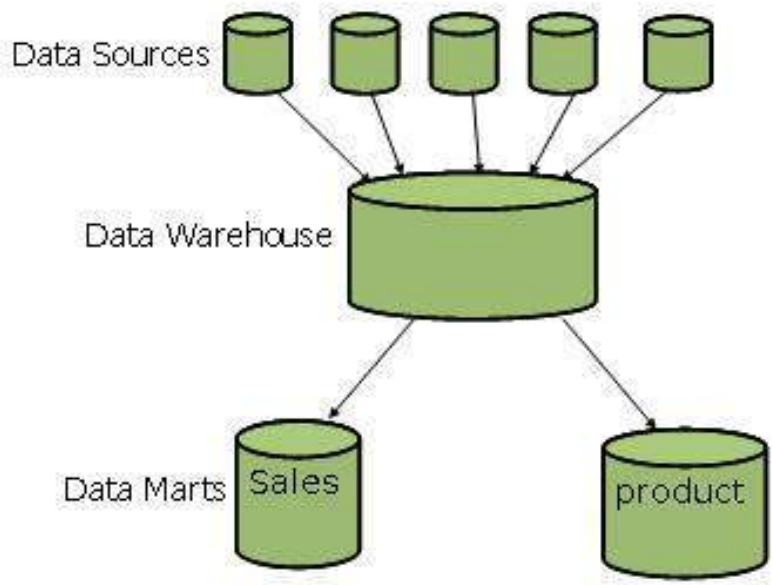
Data Mart

- Data marts contain a subset of organization-wide data that is valuable to specific groups of people in an organization.
- In other words, a data mart contains only those data that is specific to a particular group.
    - For example, the marketing data mart may contain only data related to items, customers, and sales. Data marts are confined to subjects.
- Data mart contains the subset of organization-wide data. This subset of data is valuable to specific groups of an organization.
    - In other words, we can say that a data mart contains data specific to a particular group.
- There are three main types of data marts.
    - Dependent data marts are created from an enterprise DW and use it as a main source of information (it's also known as a top-down approach).
    - Independent data marts are standalone systems that function without DWs extracting information from various external and internal sources (it's also known as a top-down approach).
    - Hybrid data marts combine information from both DW and other operational systems.

Points to Remember About Data Marts

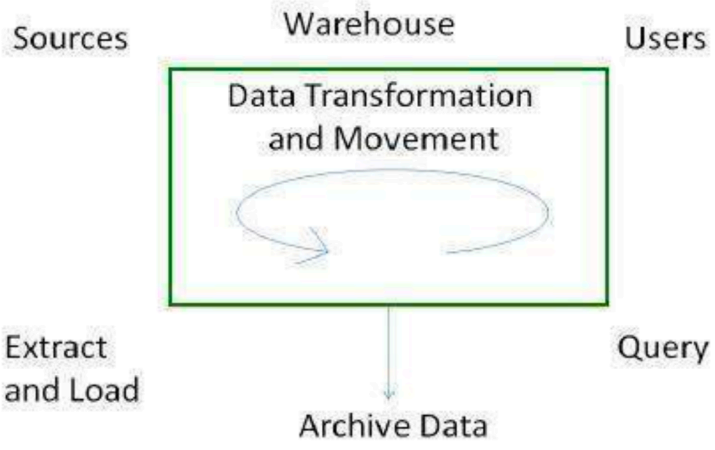
- Windows-based or Unix/Linux-based servers are used to implement data marts. They are implemented on low-cost servers.
- The implementation cycle of a data mart is measured in short periods of time, i.e., in weeks rather than months or years.
- The life cycle of data marts may be complex in the long run, if their planning and design are not organization-wide.
- Data marts are small in size.

- Data marts are customized by department.
- The source of a data mart is departmentally structured data warehouse. Data marts are flexible.



Virtual Warehouse

- The view over an operational data warehouse is known as virtual warehouse.
- It is easy to build a virtual warehouse. Building a virtual warehouse requires excess capacity on operational database servers.



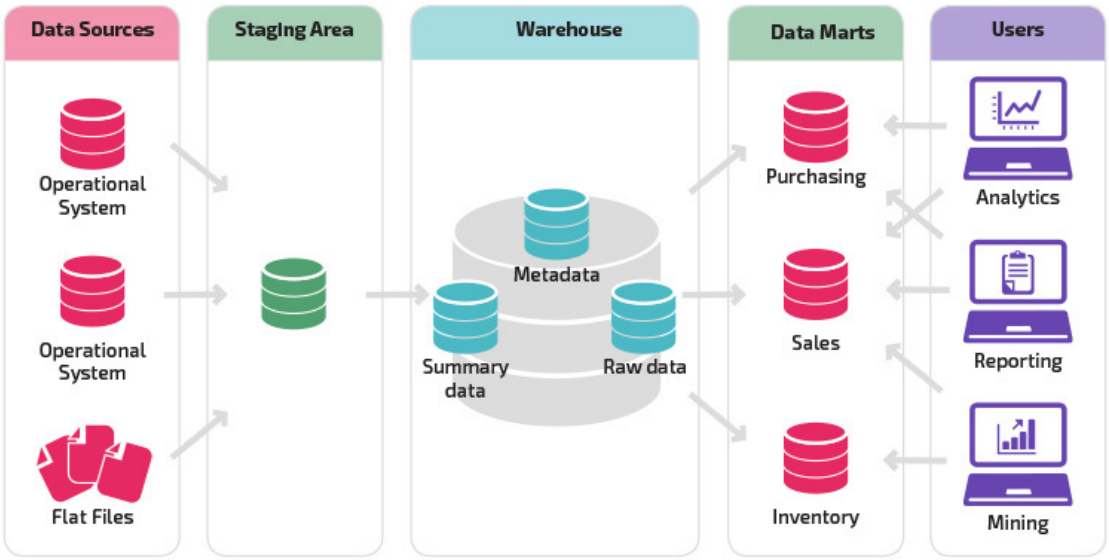
- Process Flow in Data Warehouse
- There are four major processes that contribute to a data warehouse:

1.Extract and load the data.

2.Cleaning and transforming the data.

3. Backup and archive the data.

4. Managing queries and directing them to the appropriate data sources.



What is Metadata?

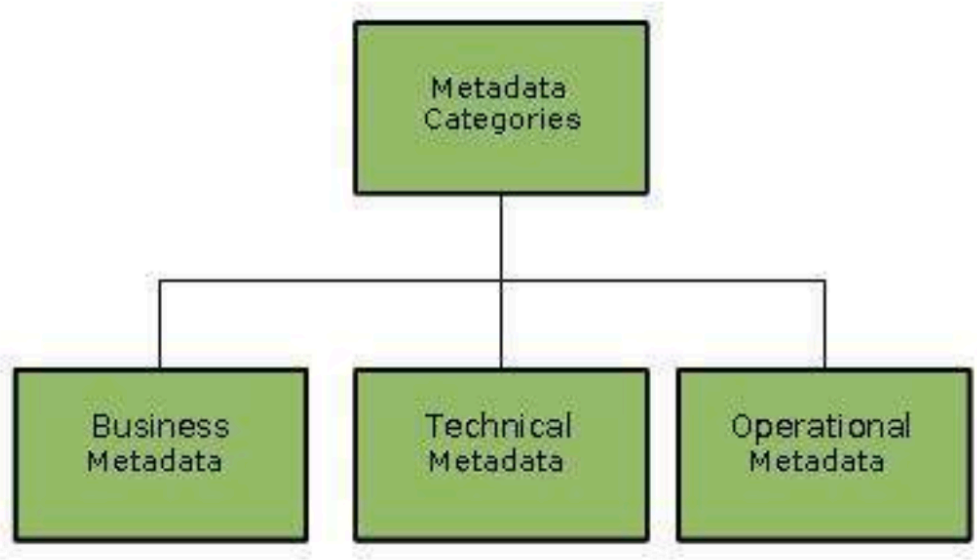
- Metadata is simply defined as data about data. The data that are used to represent other data is known as metadata.
    - For example, the index of a book serves as a metadata for the contents in the book. In other words, we can say that metadata is the summarized data that leads us to the detailed data.
- In terms of data warehouse, we can define metadata as following:
    - Metadata is a road-map to data warehouse.
    - Metadata in data warehouse defines the warehouse objects.
    - Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.

Metadata Repository

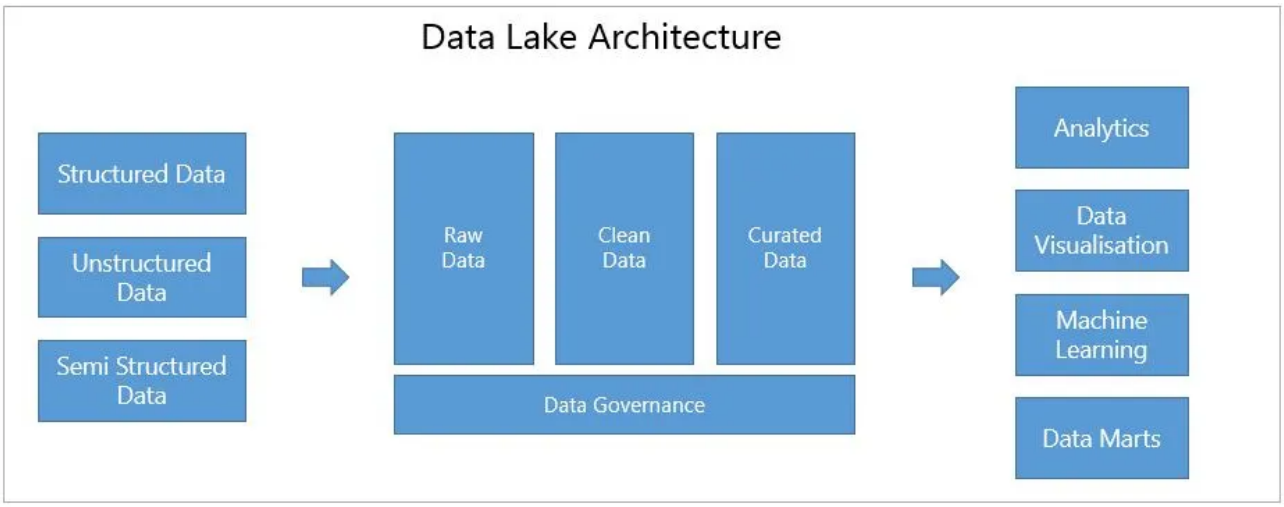
- Metadata repository is an integral part of a data warehouse system. It contains the following metadata:
    - **Business metadata** - It contains the data ownership information, business definition, and changing policies.
    - **Operational metadata** - It includes currency of data and data lineage. Currency of data refers to the data being active, archived,

or purged. Lineage of data means history of data migrated and transformation applied on it.

- **Data for mapping from operational environment to data warehouse** - It metadata includes source databases and their contents, data extraction, data partition, cleaning, transformation rules, data refresh and purging rules.
- **The algorithms for summarization** - It includes dimension algorithms, data on granularity, aggregation, summarizing, etc.



Data lake



- A Data lake is a vast pool for saving data in its native, unprocessed form. A data lake stands out for its high agility as it isn't limited to a warehouse's fixed configuration.
- A data lake is a centralized repository designed to store, process, and secure large amounts of structured, semistructured, and unstructured data. It can store data in its native format and process any variety of it, ignoring size limits
- a data lake uses the ELT approach swapping transform and load operations. Supporting large storage and scalable computing, a data lake starts data loading immediately after extracting it, handling raw — often unstructured — data.
- LT is a more advanced method as it allows for significantly increasing volumes of data to be processed. It also expedites information processing (since transformation happens only on-demand) and requires less maintenance.
- A data lake is worth building in those projects that are going to scale and would need a more advanced architecture. Besides, they are very convenient, for instance, when the purpose of data hasn't been determined yet since you can load data quickly, store it, and then modify it as necessary. Once you need data, you can apply such data processing tools as Apache or MapReduce to transform it during retrieval and analysis.
- Data lakes are also a powerful tool for data scientists and ML engineers, who would use raw data to prepare it for predictive analytics and machine learning.

Data Lakehouse

- A lakehouse is an architectural pattern that combines the best elements of data warehouses and data lakes.
- Lake houses enable you to query data across your data warehouse, data lake, and operational databases to gain faster and deeper insights that are not possible otherwise.
- With a lake house architecture, you can store data in open file formats in your data lake and query it in place while joining with data warehouse data.

- This enables you to make this data easily available to other analytics and machine learning tools, rather than locking it in a new silo.

OLAP(online analytical processing)

• An OLAP system by contrast will typically run a lower volume of transactions, but they could be longer running queries.

• An OLAP system might also run these queries over aggregated historical data, which could have been sourced from other OLTP systems, even multiple ones.

• ETL frameworks will move data from OLTP system to OLAP system.

- https://cloud.google.com/architecture/build-a-data-lake-on-gcp
- https://cloud.google.com/learn/what-is-a-data-lake
- https://cloud.google.com/solutions/data-lake

Types of OLAP Servers

We have four types of OLAP servers:

- Relational OLAP
- *ROLAP* Multidimensional OLAP
- *MOLAP* Hybrid OLAP
- *HOLAP* Specialized SQL Servers

OLTP(online Transactional processing)

- An OLTP system usually processes a high number of short transactions like select or insert statements in basic SQL terms. Queries should be fast and the system should maintain a high level of data integrity.
    - Example purchase transaction in SQL database is an OLTP system. It will modify the database.
- OLTP and OLAP are both online database systems.

Difference between OLAP and OLTP

Sr.No.	Data Warehouse *OLAP*	Operational Database*OLTP*
1	It involves historical processing of information.	It involves day-to-day processing.
2	OLAP systems are used by knowledge workers such as executives, managers, and analysts.	OLTP systems are used by clerks, DBAs, or database professionals.
3	It is used to analyze the business.	It is used to run the business.
4	It focuses on Information out.	It focuses on Data in.
5	It is based on Star Schema, Snowflake Schema, and Fact Constellation Schema.	It is based on Entity Relationship Model.
6	It focuses on Information out.	It is application oriented.
7	It contains historical data.	It contains current data.
8	It provides summarized and consolidated data.	It provides primitive and highly detailed data.
9	It provides summarized and multidimensional view of data.	It provides detailed and flat relational view of data.
10	The number of users is in hundreds.	The number of users is in thousands.
11	The number of records accessed is in millions.	The number of records accessed is in tens.
12	The database size is from 100GB to 100 TB.	The database size is from 100 MB to 100 GB.
13	These are highly flexible.	It provides high performance.

Difference between Data Warehouse and Data lake

	Data lake	Data warehouse
Type	Structured, semi-structured, unstructured Relational, non-relational	Structured Relational
Schema	Schema on read	Schema on write
Format	Raw, unfiltered	Processed, vetted
Sources	Big data, IoT, social media, streaming data	Application, business, transactional data, batch reporting
Scalability	Easy to scale at a low cost	Difficult and expensive to scale
Users	Data scientists, data engineers	Data warehouse professionals, business analysts
Use cases	Machine learning, predictive analytics, real-time analytics	Core reporting, BI