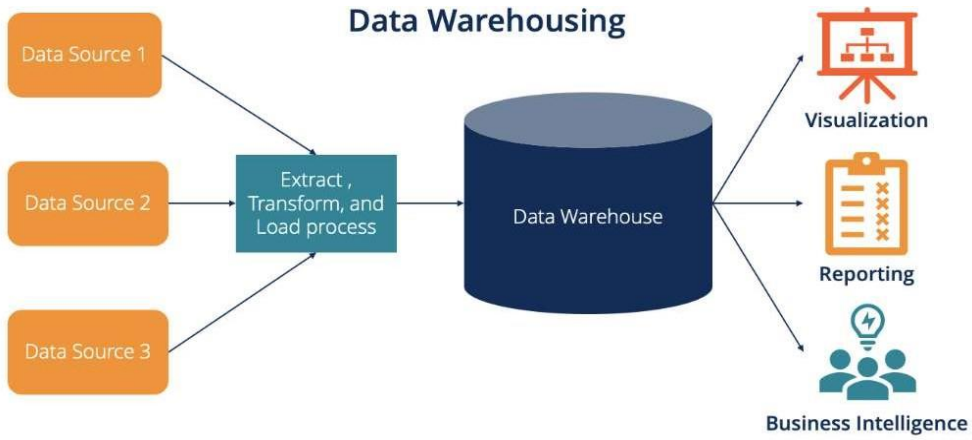
UNIT-II

What Is a Data Warehouse

Data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site. Data warehouses are constructed via a process of data cleaning, data integration, data transformation, data loading, and periodic data refreshing.

Data warehousing can be defined as the process of data collection and storage from various sources and managing it to provide valuable business insights.



Important Features of Data Warehouse

1. Subject Oriented

A data warehouse is subject-oriented. It provides useful data about a subject instead of the company's ongoing operations, and these subjects can be customers, suppliers, marketing, product, promotion, etc. A data warehouse usually focuses on modeling and analysis of data that helps the business organization to make data-driven decisions.

2. Time-Variant:

The different data present in the data warehouse provides information for a specific period.

3. Integrated

A data warehouse is built by joining data from heterogeneous sources, such as social databases, level documents, etc.

4. Non-Volatile

It means, once data entered into the warehouse cannot be change.

The following steps are involved in the process of data warehousing:

1. **Extraction of data** – A large amount of data is gathered from various sources.
2. **Cleaning of data** – Once the data is compiled, it goes through a cleaning process. The data is scanned for errors, and any error found is either corrected or excluded.
3. **Transform / Conversion of data** – After being cleaned, the format is changed from the database to a warehouse format.
4. **Storing in a warehouse** – Once converted to the warehouse format, the data stored in a warehouse goes through processes such as consolidation and summarization to make it easier and more coordinated to use. As sources get updated over time, more data is added to the warehouse.

Advantages of Data Warehouse

- More accurate data access
- Improved productivity and performance
- Cost-efficient
- Consistent and quality data

Data Warehouse Architecture

- **Source Layer**: This is where data originates. It can include databases, flat files, APIs, and other data sources.
- **Staging Area**: A temporary storage area where data is cleaned, transformed, and loaded before moving into the data warehouse.
- **Data Storage Layer**: The core of the data warehouse where cleaned and integrated data is stored. This can be organized in different ways, such as star schema, snowflake schema, or a normalized schema.
- **Data Presentation Layer**: The layer that provides data to users for querying and reporting. This can include data marts or OLAP (Online Analytical Processing) cubes.

2. ETL (Extract, Transform, Load)

- **Extract**: Collecting data from different sources.
- **Transform**: Cleaning and converting the data into a suitable format.

- **Load**: Storing the transformed data into the data warehouse.

3. Schemas

- **Star Schema**: A simple schema where a central fact table is connected to multiple dimension tables.
- **Snowflake Schema**: A more complex schema where dimension tables are normalized into multiple related tables.
- **Fact Constellation Schema**: Multiple fact tables sharing dimension tables, also known as galaxy schema.

4. Fact and Dimension Tables

- **Fact Table**: Contains quantitative data for analysis and is often denormalized.
- **Dimension Table**: Contains descriptive attributes related to the facts, providing context to the data.

5. OLAP (Online Analytical Processing)

- Enables fast querying and reporting, often using multidimensional data structures like cubes.
- **ROLAP (Relational OLAP)**: Uses relational databases to store and manage warehouse data.
- **MOLAP (Multidimensional OLAP)**: Uses multidimensional data storage.
- **HOLAP (Hybrid OLAP)**: Combines ROLAP and MOLAP.

6. Data Marts

- Subsets of data warehouses focused on specific business areas or departments. They can be dependent (sourced from a central data warehouse) or independent.

7. Data Lake

- A storage repository that holds a vast amount of raw data in its native format until it is needed.

8. Data Governance

- Policies and procedures to ensure data quality, security, and compliance.

9. Metadata

- Data about data, providing information on data sources, transformations, storage, and usage.

10. Business Intelligence Tools

- Software tools that facilitate querying, reporting, and data analysis. Examples include Tableau, Power BI, and Looker.

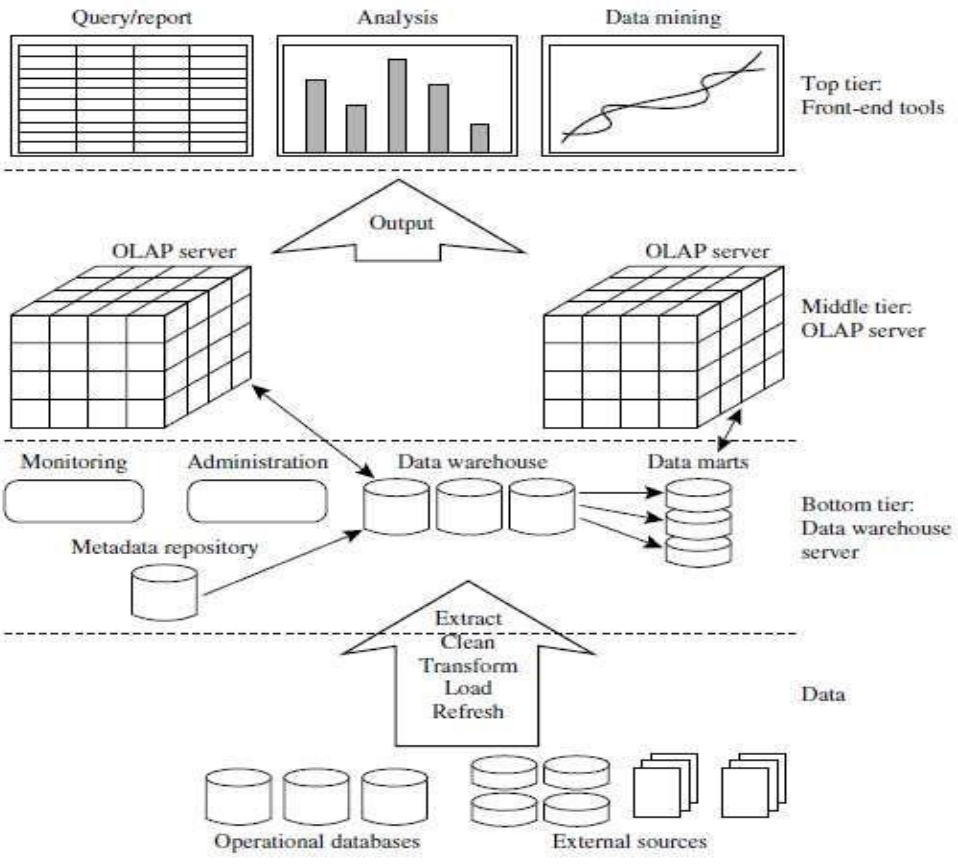
11. Performance Optimization

- Techniques like indexing, partitioning, and parallel processing to enhance the performance of the data warehouse.

12. Cloud Data Warehousing

- Data warehousing solutions offered by cloud providers such as Amazon Redshift, Google BigQuery, and Snowflake, offering scalability and flexibility.

Data warehouse Architecture



ure 4.1 A three-tier data warehousing architecture.

OLAP vs OLTP

OLAP	OLTP
Consists of historical data from various Databases.	Consists only operational current data.
It is subject oriented. Used for Data Mining, Analytics, Decision making,etc.	It is application oriented. Used for business tasks.
The data is used in planning, problem solving and decision making.	The data is used to perform day to day fundamental operations.
It provides a multi-dimensional view of different business tasks.	. It reveals a snapshot of present business tasks
Large amount of data is stored typically in TB, PB	The size of the data is relatively small as the historical data is archived. For ex MB, GB
Relatively slow as the amount of data involved is large. Queries may take hours.	Very Fast as the queries operate on 5% of the data.
It only need backup from time to time as compared to OLTP.	Backup and recovery process is maintained religiously
This data is generally managed by CEO, MD, GM.	This data is managed by clerks, managers.
Only read and rarely write operation.	Both read and write operations.

Data Warehouse Implementation

- **Requirements analysis and capacity planning:** The first process in data warehousing involves defining enterprise needs, defining architectures, carrying out capacity planning, and selecting the hardware and software tools.
- **Hardware integration:** Once the hardware and software has been selected, they require to be put by integrating the servers, the storage methods, and the user software tools.
- **Modeling:** Modeling is a significant stage that involves designing the warehouse schema and views. This may contain using a modeling tool if the data warehouses are sophisticated
- **Physical modeling:** For the data warehouses to perform efficiently, physical modeling is needed. This contains designing the physical data warehouse organization, data placement, data partitioning, deciding on access techniques, and indexing.
- **Sources:** The information for the data warehouse is likely to come from several data sources. This step contains identifying and connecting the sources using the gateway, ODBC drives, or another wrapper.
- **ETL(Extract,transform,load):** The data from the source system will require to go through an ETL phase. The process of designing and implementing the ETL phase may contain defining a suitable ETL tool vendors and purchasing and implementing the tools
- **Populate the data warehouses:** Once the ETL tools have been agreed upon, testing the tools will be needed, perhaps using a staging area. Once everything is working

adequately, the ETL tools may be used in populating the warehouses given the schema and view definition.

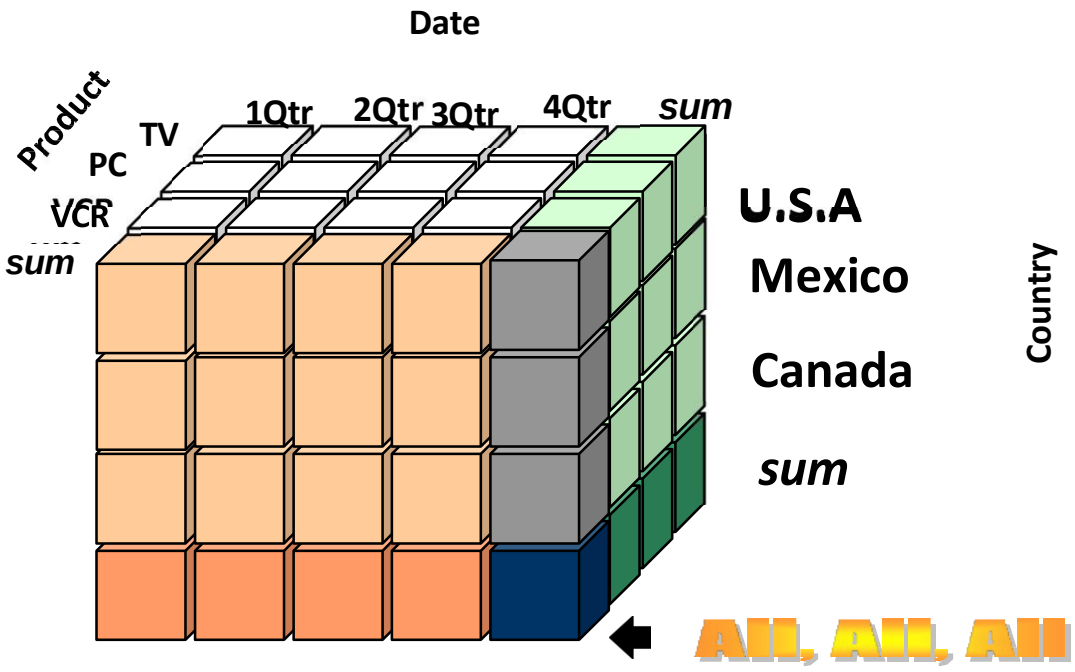
- **User applications:** For the data warehouses to be helpful, there must be end-user applications. This step contains designing and implementing applications required by the end-users.
- **Roll-out the warehouses and applications:** Once the data warehouse has been populated and the end-client applications tested, the warehouse system and the operations may be rolled out for the user's community to use.

➢ **Data ware house**

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

Data Cube: A Multidimensional Data Model

- *A **data cube allows data to be modeled and viewed in multiple*** dimensions. It is defined by dimensions and facts.
- When data is grouped or combined in multidimensional matrices called Data Cubes. The data cube method has a few alternative names or a few variants, such as "Multidimensional databases," "materialized views," and "OLAP (On-Line Analytical Processing)."
- Dimensions are the perspectives or entities with respect to which an organization wants to keep records.

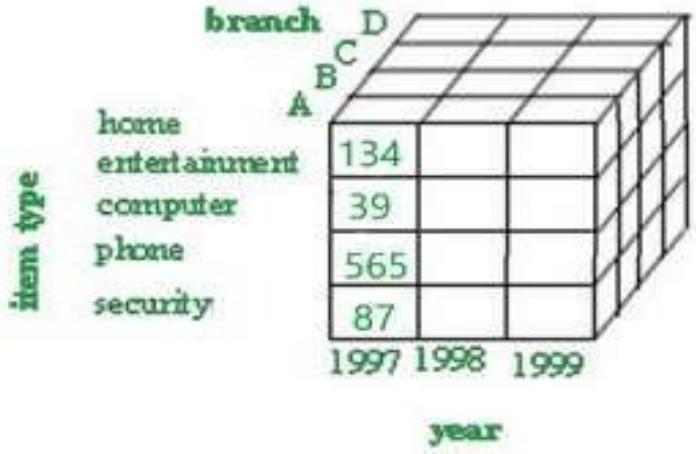


2-D view of Sales Data

location ="Vancouver"				
	item (type)			
time (quarter)	home entertainment	computer	phone	security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q3	927	1038	38	580

3-D view of Sales Data

location ="Chicago"					location ="New York"				location ="Toronto"			
item					item				item			
time	home ent.	comp.	phone	sec.	time	comp.	phone	sec.	home ent.	comp.	phone	sec.
Q1	854	882	89	623	1087	968	38	872	818	746	43	591
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784



Typical OLAP Operations

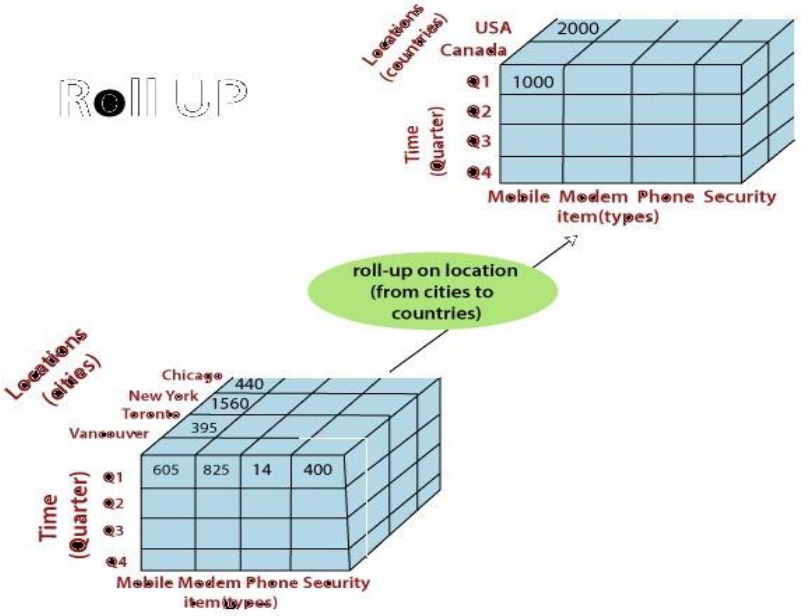
- **Rollup (drill up)**
- **drill down**

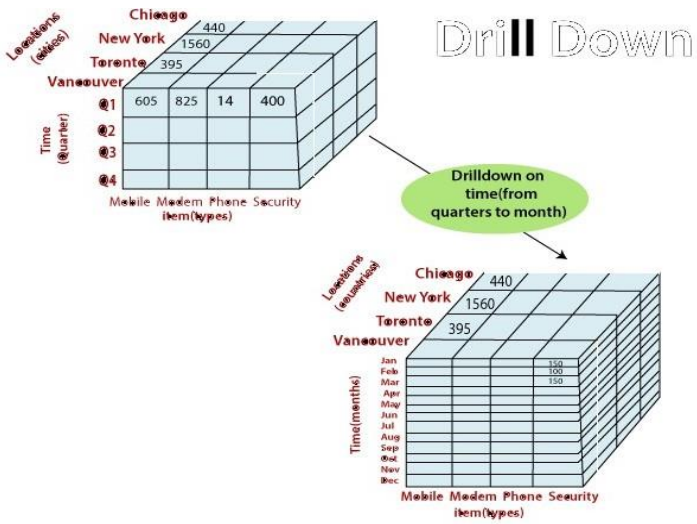
- **slice and dice**
- **pivot**
- **Roll-up:** The roll-up operation (also called the *drill-up operation by some vendors)* performs aggregation on a data cube, either by *climbing up a concept hierarchy for* a dimension or by *dimension reduction.*
- **Drill-down:** Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either *stepping down a concept hierarchy for a dimension or introducing additional dimensions*

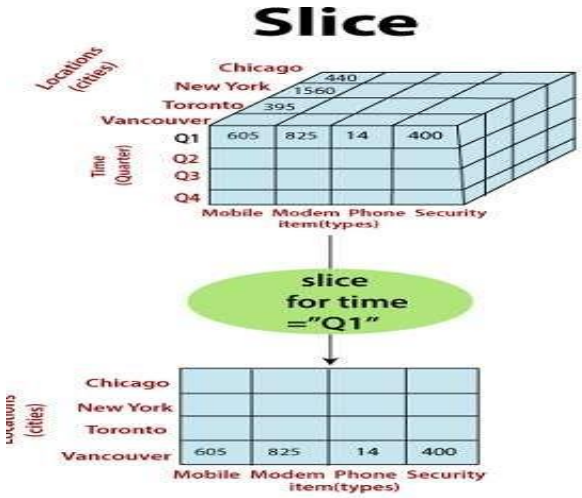
The **slice operation** performs a selection on one dimension of the given cube, resulting in a sub-cube. Reduces the dimensionality of the cubes.

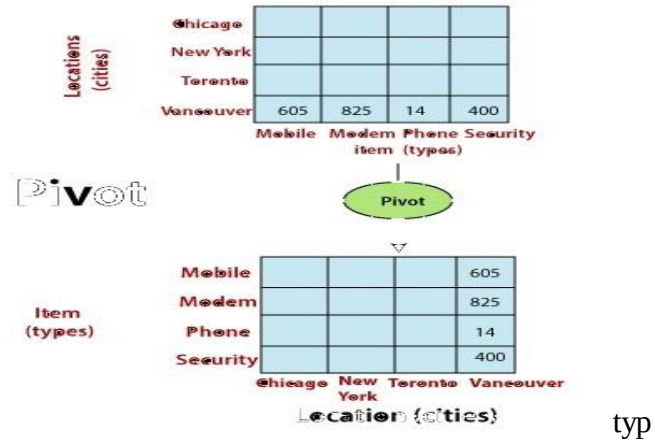
The **dice operation** defines a sub-cube by performing a selection on two or more dimensions.

- **Pivot**
- Pivot is also known as rotate. It Rotates the data axis to view the data from different perspectives.

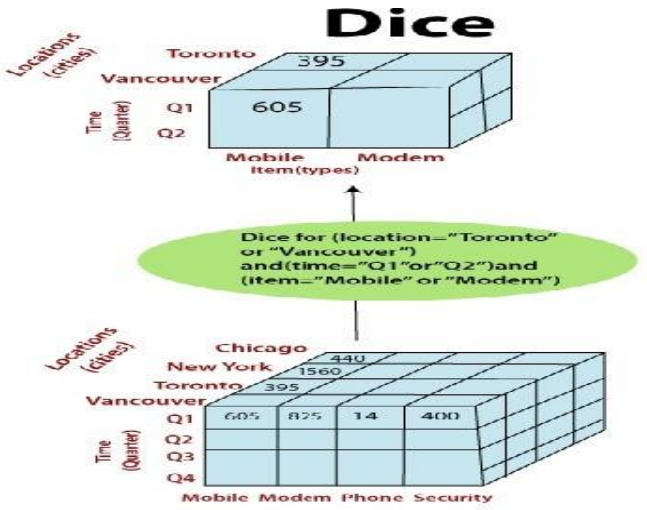



Drill Down
Chicago
New York
Toronto
Vancouver
440
1560
395
605
825
14
400
Q1
Q2
Q3
Q4
Time (Quarter)
Mobile Modem Phone Security
item(types)
Drilldown on time(from quarters to month)
Time(months)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
150
100
150


Slice
Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
605
825
14
400
Q1
Q2
Q3
Q4
Time (Quarter)
Mobile Modem Phone Security
item(types)
slice for time ="Q1"



typ



Types of OLAP servers

ROLAP

MOLAP

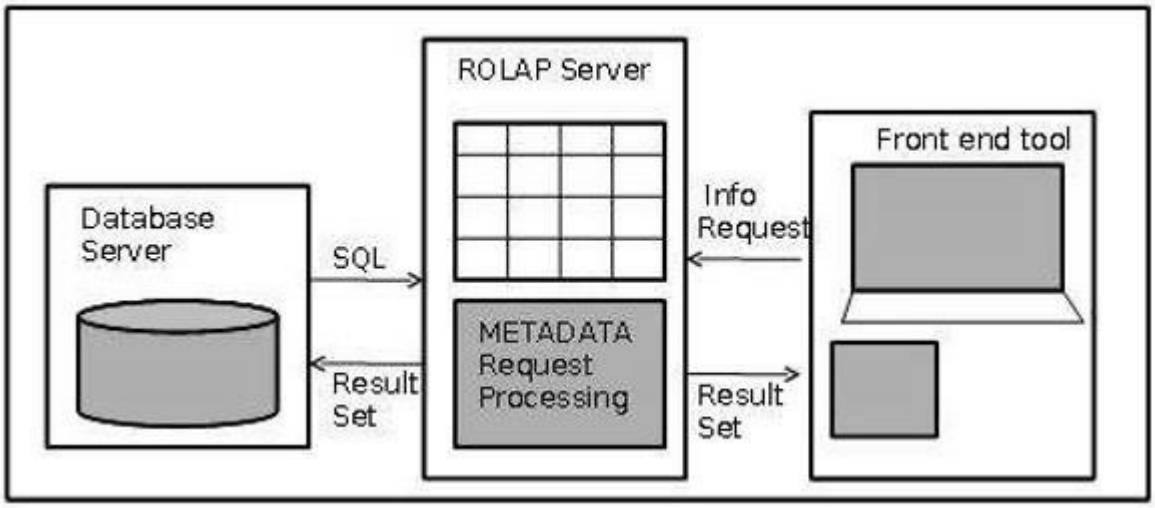
HOLAP

ROLAP

ROLAP stands for Relational Online Analytical Processing. ROLAP stores data in columns and rows (also known as relational tables) and retrieves the information on demand through user submitted queries. A ROLAP database can be accessed through complex SQL queries to calculate information.

ROLAP includes the following components –

- Database server
- ROLAP server
- Front-end tool.



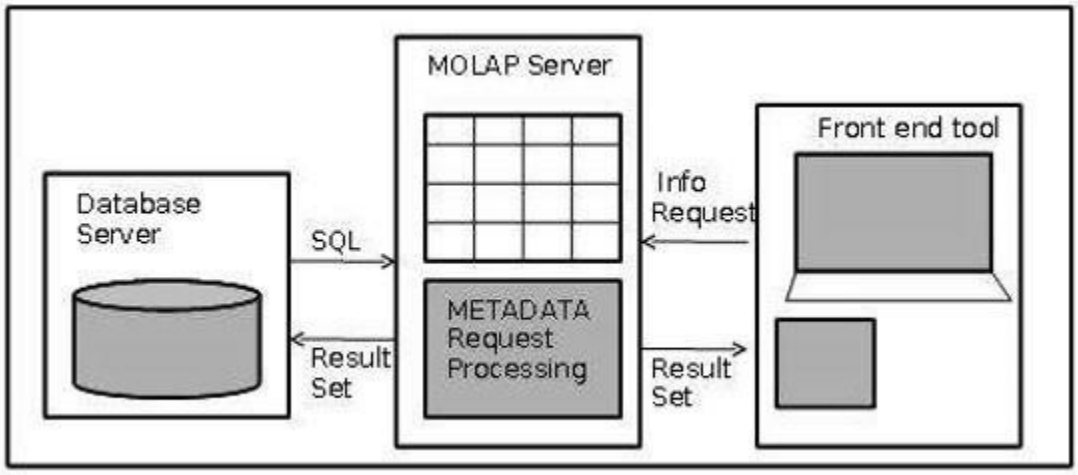
MOLAP

Multidimensional OLAP (MOLAP) uses array-based multidimensional storage engines for multidimensional views of data. With multidimensional data stores, the storage utilization may be low if the dataset is sparse.

Uses multidimensional data cubes for data storage, providing fast query performance.

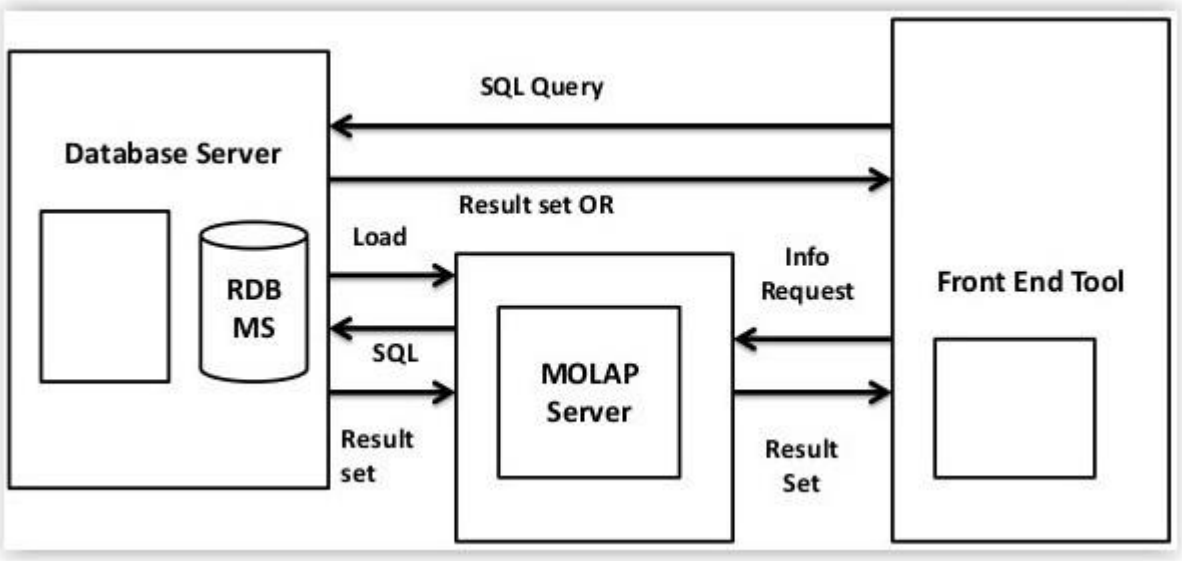
MOLAP includes the following components –

- Database server.
- MOLAP server.
- Front-end tool.



HOLAP

HOLAP stands for Hybrid Online Analytical Processing. As the name suggests, the HOLAP storage mode connects attributes of both MOLAP and ROLAP. Since HOLAP involves storing part of your data in a ROLAP store and another part in a MOLAP store, developers get the benefits of both.



MOLAP vs ROLAP

Sr.No.	MOLAP	ROLAP
1	Information retrieval is fast.	Information retrieval is comparatively slow.
2	Uses sparse array to store data-sets.	Uses relational table.
3	MOLAP is best suited for inexperienced users, since it is very easy to use.	ROLAP is best suited for experienced users.
4	Maintains a separate database for data cubes.	It may not require space other than available in the Data warehouse.
5	DBMS facility is weak.	DBMS facility is strong.

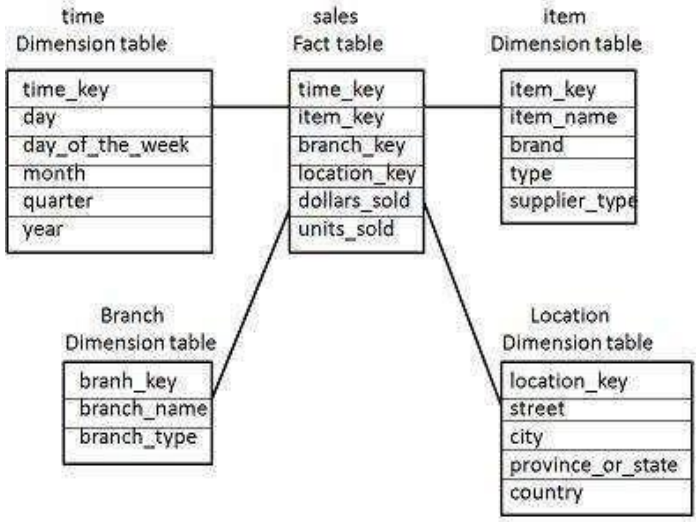
Data Mining	Data Warehousing
Data mining is the process of determining data patterns.	A data warehouse is a database system designed for analytics.
Data mining is generally considered as the process of extracting useful data from a large set of data.	Data warehousing is the process of combining all the relevant data.
Business entrepreneurs carry data mining with the help of engineers.	Data warehousing is entirely carried out by the engineers.
In data mining, data is analyzed repeatedly.	In data warehousing, data is stored periodically.
Data mining uses pattern recognition techniques to identify patterns.	Data warehousing is the process of extracting and storing data that allow easier reporting.
One of the most amazing data mining technique is the detection and identification of the unwanted errors that occur in the system.	One of the advantages of the data warehouse is its ability to update frequently. That is the reason why it is ideal for business entrepreneurs who want up to date with the latest stuff.
The data mining techniques are cost-efficient as compared to other statistical data applications.	The responsibility of the data warehouse is to simplify every type of business data.
The data mining techniques are not 100 percent accurate. It may lead to serious consequences in a certain condition.	In the data warehouse, there is a high possibility that the data required for analysis by the company may not be integrated into the warehouse. It can simply lead to loss of data.
Companies can benefit from this analytical tool by equipping suitable and accessible knowledge-based data.	Data warehouse stores a huge amount of historical data that helps users to analyze different periods and trends to make future predictions.

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

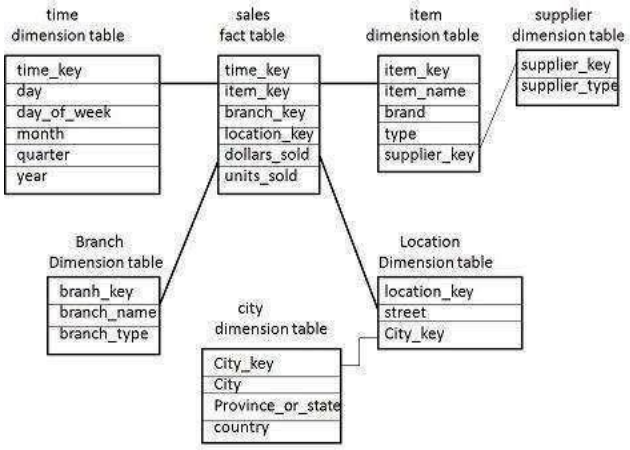
The most popular data model for a data warehouse is a **multidimensional model,** which can exist in the form of a **star schema, a snowflake schema, or a fact constellation schema.**

Star schema: The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (**fact table)** containing the bulk of the data, with no redundancy, (2) a set of smaller attendant tables (**dimension tables),** one for each dimension**.**

- Star schema of *sales data warehouse*

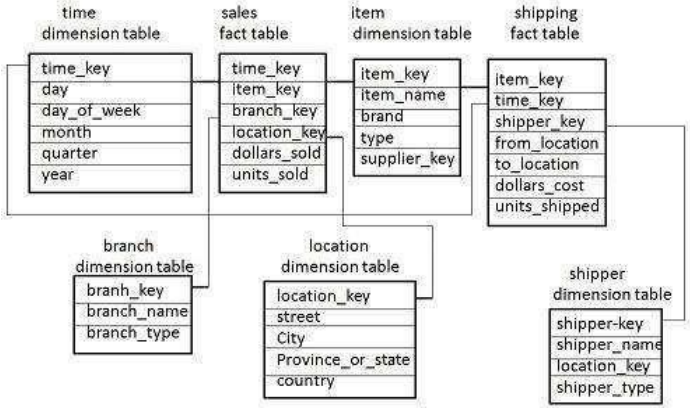


- **Snowflake schema:**
- The snowflake schema is a variant of the star schema model, where some dimension tables are *normalized, thereby further splitting the data into* additional tables. The resulting schema graph forms a shape similar to a snowflake
- The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space.



- **Fact Constellation Schema**

A fact constellation has multiple fact tables. It is also known as galaxy schema. The following diagram shows two fact tables, namely sales and shipping



- The sales fact table is same as that in the star schema.
- The shipping fact table has the five dimensions, namely item_key, time_key, shipper_key, from_location, to_location.
- The shipping fact table also contains two measures, namely dollars sold and units sold.
- It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table