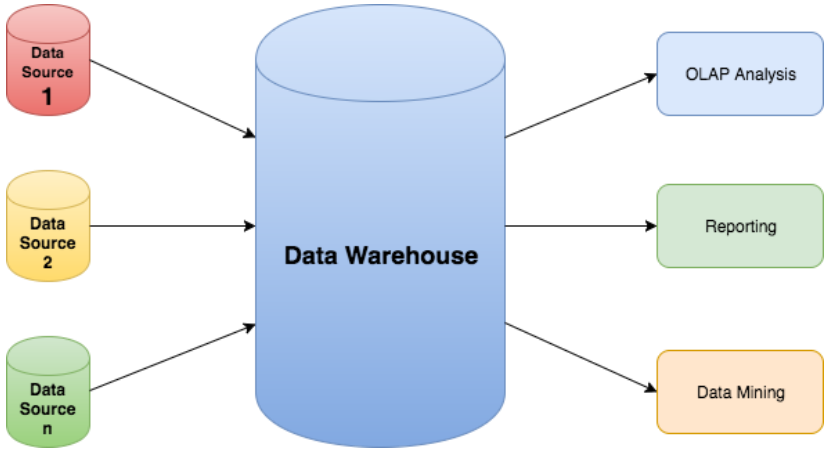
WAREHOUSING

A data warehouse is a place which store information collected from multiple sources under unified schema. Information stored in a data warehouse is critical to organizations for the process of decision-making.



Integrating Data Mining System with a Database or Data Warehouse System

The data mining system needs to be integrated with database or the data warehouse system. If the data mining system is not integrated with any database or data warehouse system then there will be no system to communicate with. This scheme is known as non-coupling scheme. In this scheme the main focus is put on data mining design and for developing efficient and effective algorithms for mining the available data sets.

Here is the list of Integration Schemes:

- **No Coupling** - In this scheme the Data Mining system does not utilize any of the database or data warehouse functions. It then fetches the data from a particular source and process that data using some data mining algorithms. The data mining result is stored in other file.
- **Loose Coupling** - In this scheme the data mining system may use some of the functions of database and data warehouse system. It then fetches the data from data respiratory managed by these systems and perform data mining on that data. It then stores the mining result either in a file or in a designated place in a database or data warehouse.
- **Semi-tight Coupling** - In this scheme the data mining system is along with the kinking the efficient implementation of data mining primitives can be provided in database or data warehouse systems.
- **Tight coupling** - In this coupling scheme data mining system is smoothly integrated into database or data warehouse system. The data mining subsystem is treated as one functional component of an information system.

Data Warehousing:

Data warehousing is the process which is used to integrate data from multiple sources and then combine it into a single databasel

Why use Data Warehouse?

Some most Important reasons for using Data warehouse are:

- Integrates many sources of data and helps to decrease stress on a production system.
- Optimized Data for reading access and consecutive disk scans.
- Data Warehouse helps to protect Data from the source system upgrades.
- Allows users to perform master Data Management.
- Improve data quality in source systems.

OLAP (online analytical processing)

OLAP (online analytical processing) is a computing method that enables users to easily and selectively extract and query data in order to analyze it from different points of view. OLAP business intelligence queries often aid in trends analysis, financial reporting, sales forecasting, budgeting and other planning purposes.

For example, a user can request that data be analyzed to display a spreadsheet showing all of a company's beach ball products sold in Florida in the month of July, compare revenue figures with those for the same products in September and then see a comparison of other product sales in Florida in the same time period.

How OLAP systems work

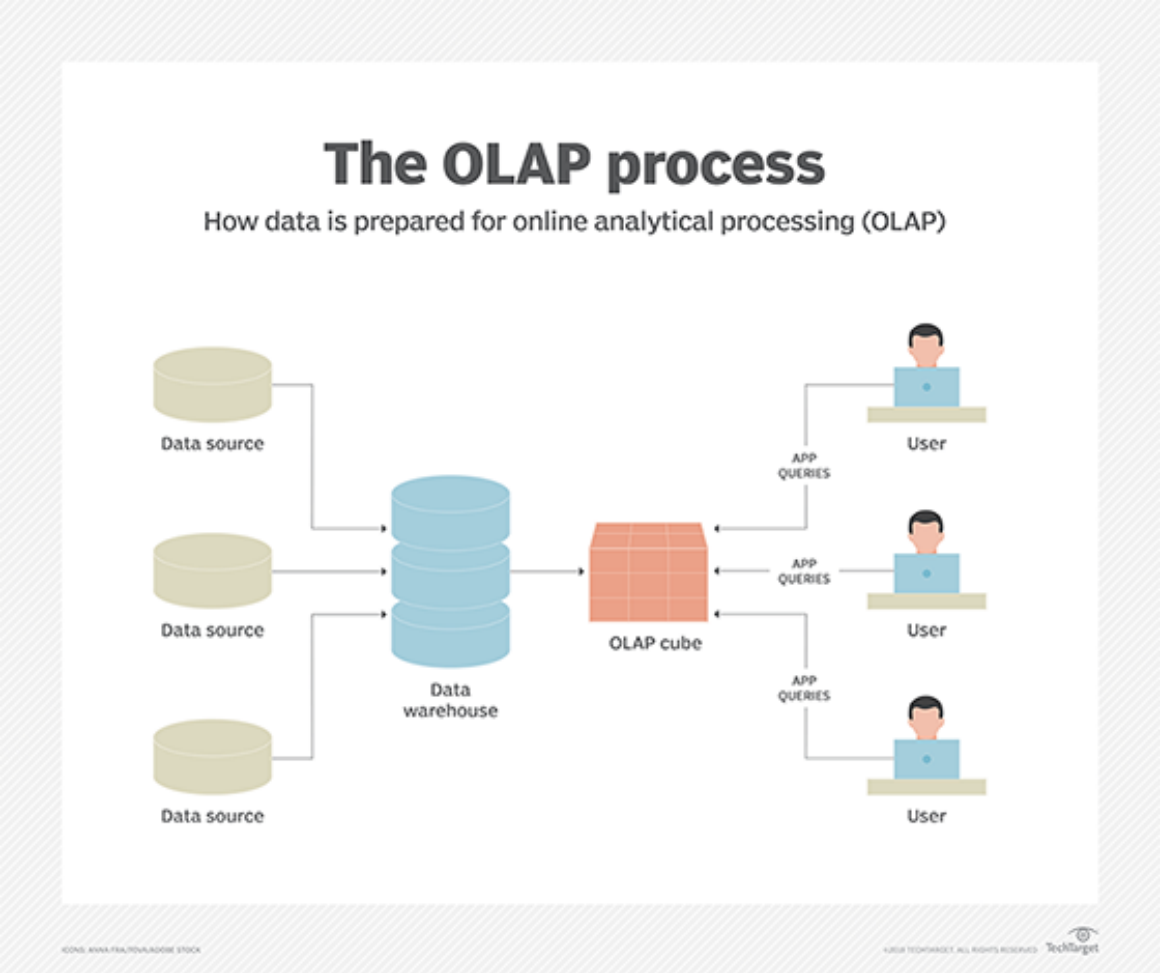
To facilitate this kind of analysis, data is collected from multiple data sources and stored in data warehouses then cleansed and organized into data cubes. Each OLAP cube contains data categorized by dimensions (such as customers, geographic sales region and time period) derived by dimensional tables in the data warehouses. Dimensions are then populated by members (such as customer names, countries and months) that are organized hierarchically. OLAP cubes are often pre-summarized across dimensions to drastically improve query time over relational databases.

Analysts can then perform five types of OLAP analytical operations against these multidimensional databases:

- **Roll-up.** Also known as *consolidation*, or *drill-up*, this operation summarizes the data along the dimension.
- **Drill-down.** This allows analysts to navigate deeper among the dimensions of data, for example drilling down from "time period" to "years" and "months" to chart sales growth for a product.

- **Slice.** This enables an analyst to take one level of information for display, such as "sales in 2017."
- **Dice.** This allows an analyst to select data from multiple dimensions to analyze, such as "sales of blue beach balls in Iowa in 2017."
- **Pivot.** Analysts can gain a new view of data by rotating the data axes of the cube.

OLAP software then locates the intersection of dimensions, such as all products sold in the Eastern region above a certain price during a certain time period, and displays them. The result is the "measure"; each OLAP cube has at least one to perhaps hundreds of measures, which are derived from information stored in fact tables in the data warehouse.



Types of OLAP systems

OLAP (online analytical processing) systems typically fall into one of three types:

Multidimensional OLAP (MOLAP) is OLAP that indexes directly into a multidimensional database.

Relational OLAP (ROLAP) is OLAP that performs dynamic multidimensional analysis of data stored in a relational database.

Hybrid OLAP (HOLAP) is a combination of ROLAP and MOLAP. HOLAP was developed to combine the greater data capacity of ROLAP with the superior processing capability of MOLAP.

Uses of OLAP

OLAP can be used for data mining or the discovery of previously undiscerned relationships between data items. An OLAP database does not need to be as large as a data warehouse, since not all transactional data is needed for trend analysis. Using Open Database Connectivity (ODBC), data can be imported from existing relational databases to create a multidimensional database for OLAP.

OLAP Operations

Since OLAP servers are based on multidimensional view of data, we will discuss OLAP operations in multidimensional data.

Here is the list of OLAP operations −

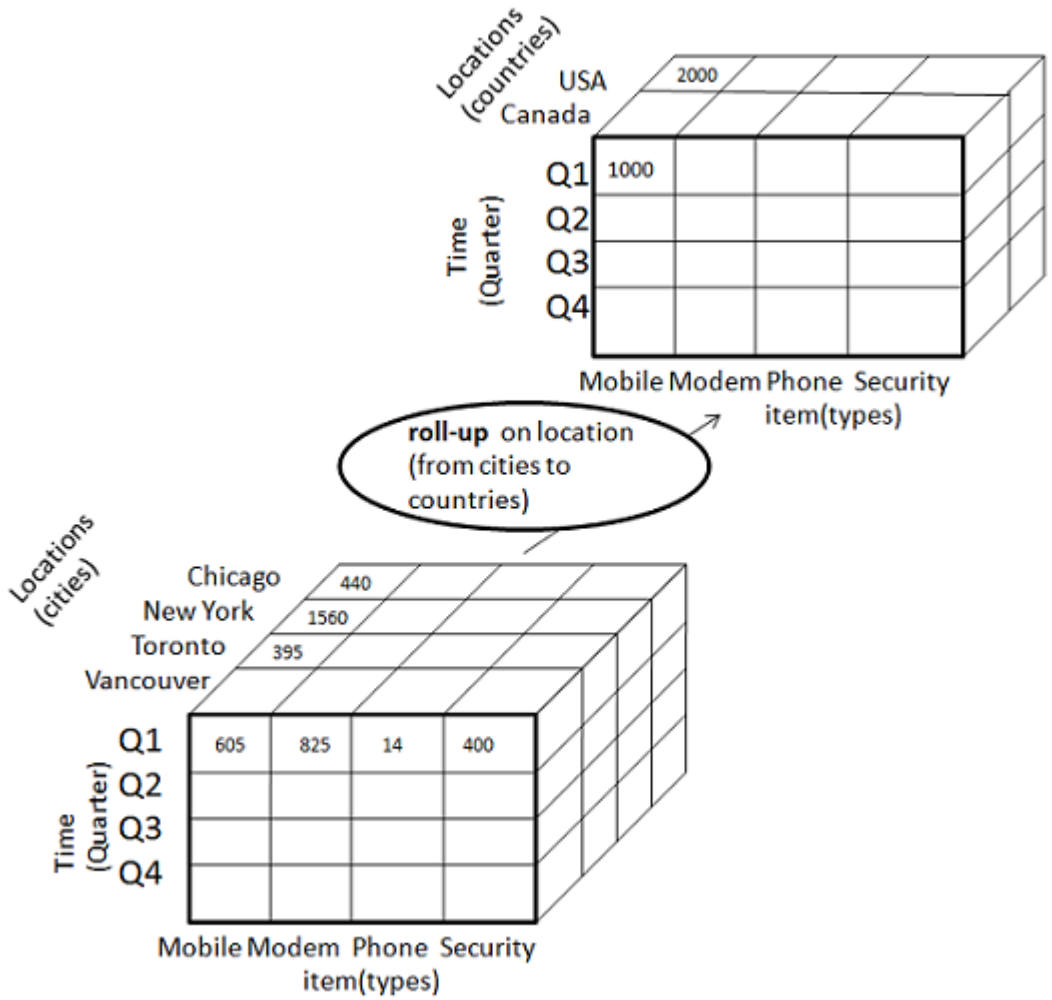
- Roll-up
- Drill-down
- Slice and dice
- Pivot (rotate)

Roll-up

Roll-up performs aggregation on a data cube in any of the following ways −

- By climbing up a concept hierarchy for a dimension
- By dimension reduction

The following diagram illustrates how roll-up works.



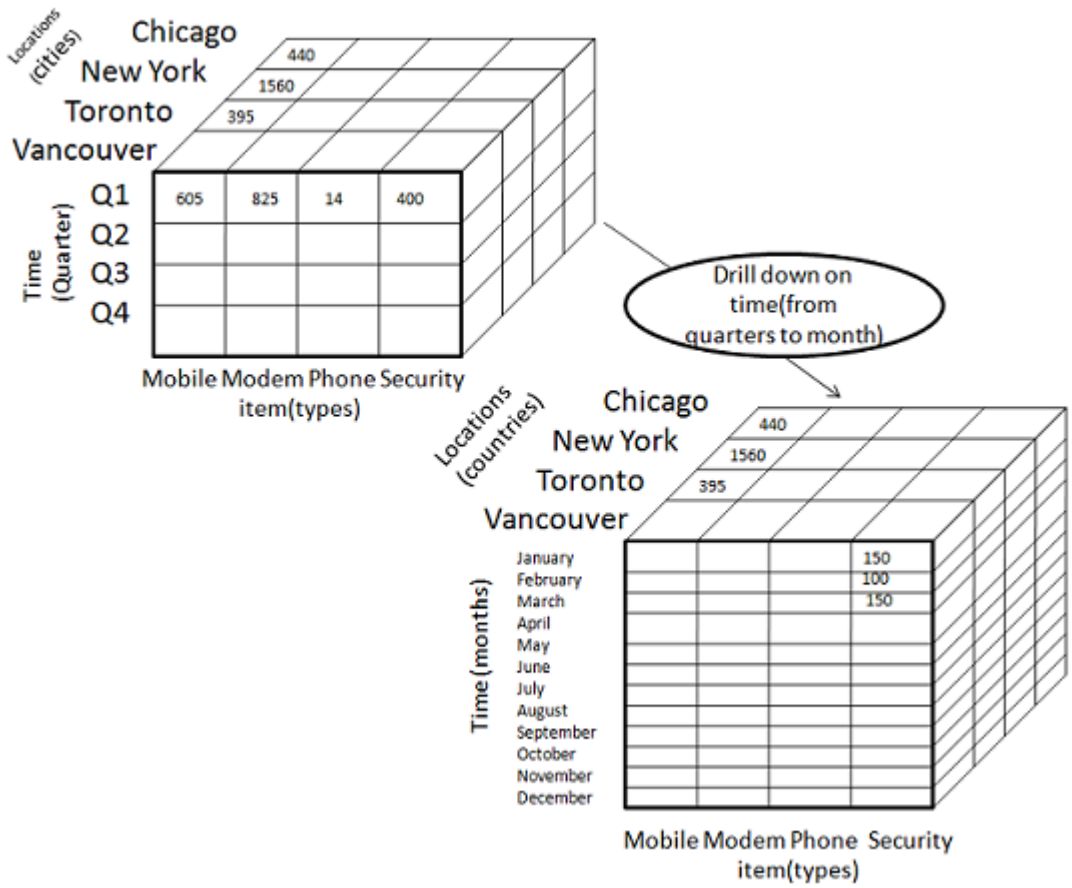
- Roll-up is performed by climbing up a concept hierarchy for the dimension location.
- Initially the concept hierarchy was "street < city < province < country".
- On rolling up, the data is aggregated by ascending the location hierarchy from the level of city to the level of country.
- The data is grouped into cities rather than countries.
- When roll-up is performed, one or more dimensions from the data cube are removed.

Drill-down

Drill-down is the reverse operation of roll-up. It is performed by either of the following ways −

- By stepping down a concept hierarchy for a dimension
- By introducing a new dimension.

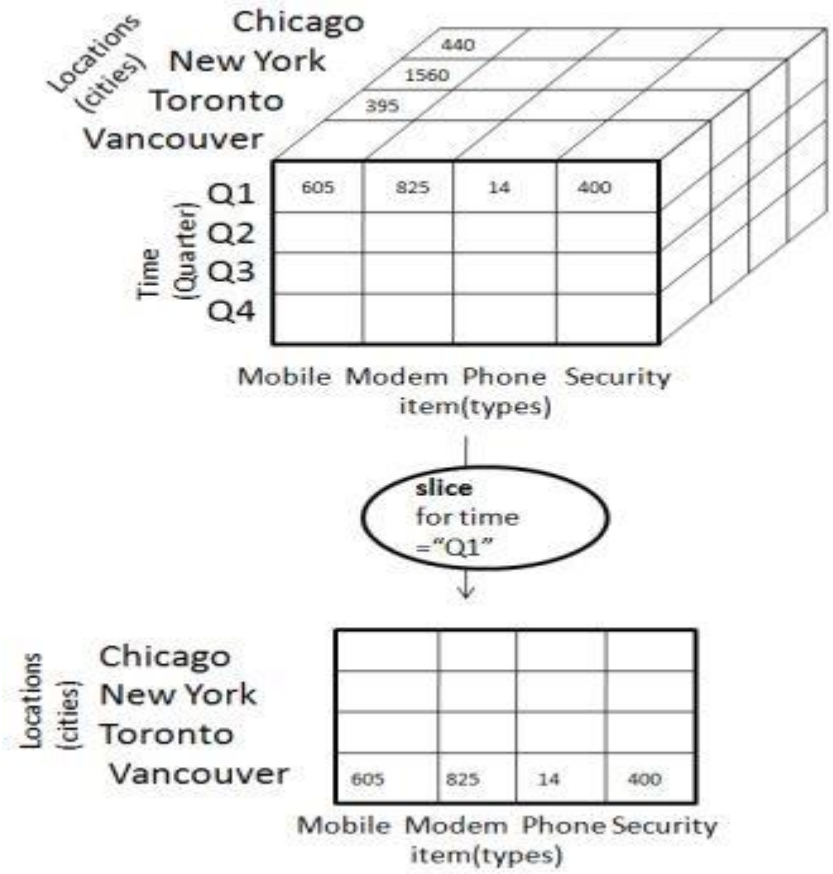
The following diagram illustrates how drill-down works −



- Drill-down is performed by stepping down a concept hierarchy for the dimension time.
- Initially the concept hierarchy was "day < month < quarter < year."
- On drilling down, the time dimension is descended from the level of quarter to the level of month.
- When drill-down is performed, one or more dimensions from the data cube are added.
- It navigates the data from less detailed data to highly detailed data.

Slice

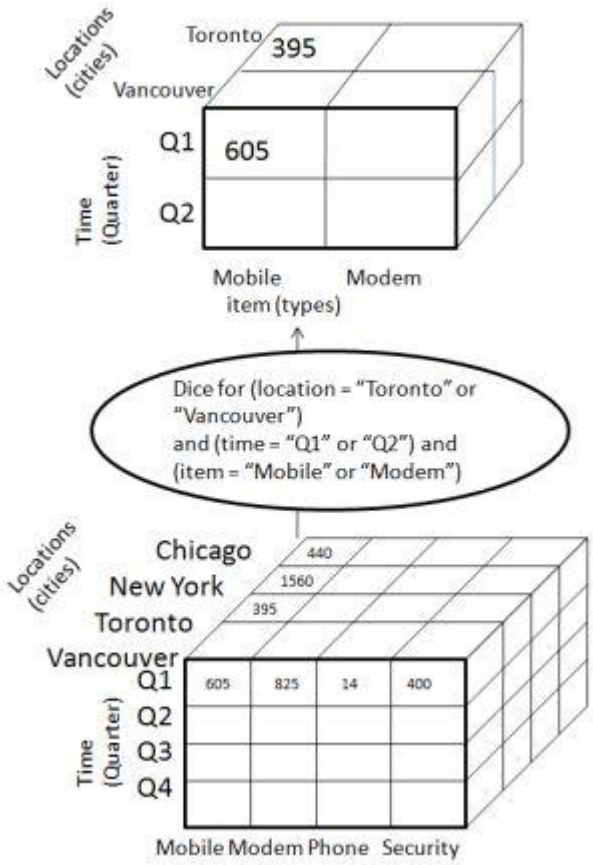
The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Consider the following diagram that shows how slice works.



- Here Slice is performed for the dimension "time" using the criterion time = "Q1".
- It will form a new sub-cube by selecting one or more dimensions.

Dice

Dice selects two or more dimensions from a given cube and provides a new sub-cube. Consider the following diagram that shows the dice operation.

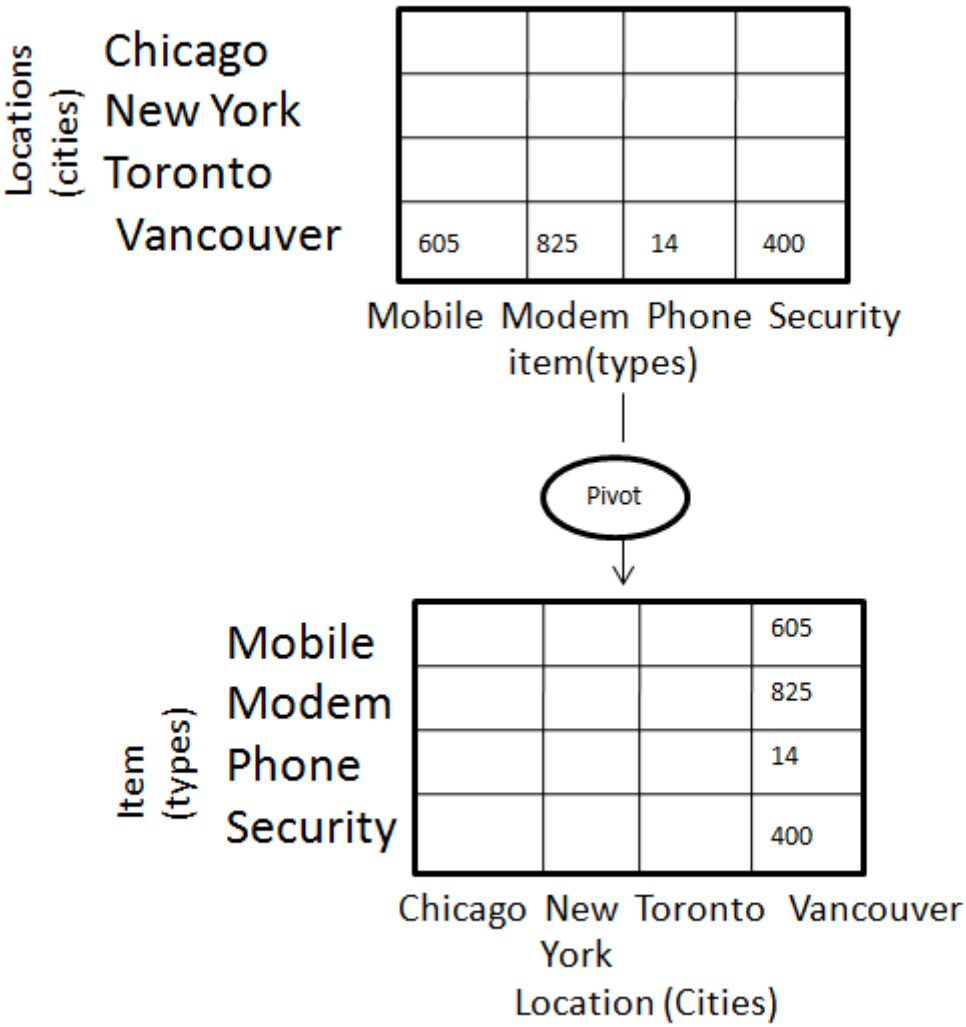


The dice operation on the cube based on the following selection criteria involves three dimensions.

- (location = "Toronto" or "Vancouver")
- (time = "Q1" or "Q2")
- (item =" Mobile" or "Modem")

Pivot

The pivot operation is also known as rotation. It rotates the data axes in view in order to provide an alternative presentation of data. Consider the following diagram that shows the pivot operation.



OLAP vs OLTP

Sr.No.	Data Warehouse (OLAP)	Operational Database (OLTP)
1	Involves historical processing of information.	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	Useful in analyzing the business.	Useful in running the business.

4	It focuses on Information out.	It focuses on Data in.
5	Based on Star Schema, Snowflake, Schema and Fact Constellation Schema.	Based on Entity Relationship Model.
6	Contains historical data.	Contains current data.
7	Provides summarized and consolidated data.	Provides primitive and highly detailed data.
8	Provides summarized and multidimensional view of data.	Provides detailed and flat relational view of data.
9	Number or users is in hundreds.	Number of users is in thousands.
10	Number of records accessed is in millions.	Number of records accessed is in tens.
11	Database size is from 100 GB to 1 TB	Database size is from 100 MB to 1 GB.
12	Highly flexible.	Provides high performance.