UNIT 1:

Data Warehousing Introduction:

What is Data Warehousing?

Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries, and decision making. Data warehousing involves data cleaning, data integration, and data consolidations.

A data warehouse is the secure electronic storage of information by a business or other organization. The goal of a data warehouse is to create a trove of historical data that can be retrieved and analyzed to provide useful insight into the organization's operations.

A data warehouse is a vital component of business intelligence. That wider term encompasses the information infrastructure that modern businesses use to track their past successes and failures and inform their decisions for the future.

KEY TAKEAWAYS

- A data warehouse is the storage of information over time by a business or other organization.
- New data is periodically added by people in various key departments such as marketing and sales.
- The warehouse becomes a library of historical data that can be retrieved and analyzed in order to inform decision-making in the business.
- The key factors in building an effective data warehouse include defining the information that is critical to the organization and identifying the sources of the information.
- A database is designed to supply real-time information. A data warehouse is designed as an archive of historical information.

Using Data Warehouse Information

There are decision support technologies that help utilize the data available in a data warehouse. These technologies help executives to use the warehouse quickly and effectively. They can gather data, analyze it, and take decisions based on the information present in the warehouse. The information gathered in a warehouse can be used in any of the following domains −

- **Tuning Production Strategies** − The product strategies can be well tuned by repositioning the products and managing the product portfolios by comparing the sales quarterly or yearly.
- **Customer Analysis** − Customer analysis is done by analyzing the customer's buying preferences, buying time, budget cycles, etc.
- **Operations Analysis** − Data warehousing also helps in customer relationship management, and making environmental corrections. The information also allows us to analyze business operations.

Integrating Heterogeneous Databases

To integrate heterogeneous databases, we have two approaches −

- Query-driven Approach
- Update-driven Approach

Query-Driven Approach

This is the traditional approach to integrate heterogeneous databases. This approach was used to build wrappers and integrators on top of multiple heterogeneous databases. These integrators are also known as mediators.

Process of Query-Driven Approach

- When a query is issued to a client side, a metadata dictionary translates the query into an appropriate form for individual heterogeneous sites involved.
- Now these queries are mapped and sent to the local query processor.
- The results from heterogeneous sites are integrated into a global answer set.

Disadvantages

- Query-driven approach needs complex integration and filtering processes.
- This approach is very inefficient.
- It is very expensive for frequent queries.
- This approach is also very expensive for queries that require aggregations.

Update-Driven Approach

This is an alternative to the traditional approach. Today's data warehouse systems follow update-driven approach rather than the traditional approach discussed earlier. In update-driven approach, the information from multiple heterogeneous sources are integrated in advance and are stored in a warehouse. This information is available for direct querying and analysis.

Advantages

This approach has the following advantages −

- This approach provide high performance.
- The data is copied, processed, integrated, annotated, summarized and restructured in semantic data store in advance.
- Query processing does not require an interface to process data at local sources.

Functions of Data Warehouse Tools and Utilities

The following are the functions of data warehouse tools and utilities −

- **Data Extraction** − Involves gathering data from multiple heterogeneous sources.
- **Data Cleaning** − Involves finding and correcting the errors in data.
- **Data Transformation** − Involves converting the data from legacy format to warehouse format.
- **Data Loading** − Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions.
- **Refreshing** − Involves updating from data sources to warehouse.

How a Data Warehouse Works

The need to warehouse data evolved as businesses began relying on computer systems to create, file, and retrieve important business documents. The concept of data warehousing was introduced in 1988 by IBM researchers Barry Devlin and Paul Murphy.1

Data warehousing is designed to enable the analysis of historical data. Comparing data consolidated from multiple heterogeneous sources can provide insight into the performance of a company. A data warehouse is designed to allow its users to run queries and analyses on historical data derived from transactional sources.

Data added to the warehouse does not change and cannot be altered. The warehouse is the source that is used to run analytics on past events, with a focus on changes over time. Warehoused data must be stored in a manner that is secure, reliable, easy to retrieve, and easy to manage.

Maintaining a Data Warehouse

There are certain steps that are taken to maintain a data warehouse. One step is data extraction, which involves gathering large amounts of data from multiple source points. After a set of data has been compiled, it goes through data cleaning, the process of combing through it for errors and correcting or excluding any that are found.

The cleaned-up data is then converted from a database format to a warehouse format. Once stored in the warehouse, the data goes through sorting, consolidating, and summarizing, so that it will be easier to use. Over time, more data is added to the warehouse as the various data sources are updated.

Data Warehouse Architecture

Designing a data warehouse is known as data warehouse architecture and depending on the needs of the data warehouse, can come in a variety of tiers. Typically there are tier one, tier two, and tier three architecture designs.

Single-tier Architecture: Single-tier architecture is hardly used in the creation of data warehouses for real-time systems. They are often used for batch and real-time processing to process operational data. A single-tier design is composed of a single layer of hardware with the goal of keeping data space at a minimum.

Two-tier Architecture: In a two-tier architecture design, the analytical process is separated from the business process. The point of this is to increase levels of control and efficiency.

Three-tier Architecture: A three-tier architecture design has a top, middle, and bottom tier; these are known as the source layer, the reconciled layer, and the data warehouse layer. This design is suited for systems with long life cycles. When changes are made in the data, an extra layer of review and analysis of the data is completed to ensure there have been no errors.

Regardless of the tier, all data warehouse architectures must meet the same five properties: separation, scalability, extensibility, security, and administrability.

Data Warehouse vs. Database

A data warehouse is not the same as a database:

- A database is a transactional system that monitors and updates real-time data in order to have only the most recent data available.
- A data warehouse is programmed to aggregate structured data over time.

For example, a database might only have the most recent address of a customer, while a data warehouse might have all the addresses of the customer for the past 10 years.

Data mining relies on the data warehouse. The data in the warehouse is sifted for insights into the business over time.

Data Warehouse vs. Data Lake

Both data warehouses and data lakes hold data for a variety of needs. The primary difference is that a data lake holds raw data of which the goal has not yet been determined. A data warehouse, on the other hand, holds refined data that has been filtered to be used for a specific purpose.

Data lakes are primarily used by data scientists while data warehouses are most often used by business professionals. Data lakes are also more easily accessible and easier to update while data warehouses are more structured and any changes are more costly.

Data Warehouse vs. Data Mart

A data mart is just a smaller version of a data warehouse. A data mart collects data from a small number of sources and focuses on one subject area. Data marts are faster and easier to use than data warehouses.

Data marts typically function as a subset of a data warehouse to focus on one area for analytical purposes, such as a specific department within an organization. Data marts are used to help make business decisions by helping with analysis and reporting.

Advantages and Disadvantages of Data Warehouses

A data warehouse is intended to give a company a competitive advantage. It creates a resource of pertinent information that can be tracked over time and analyzed in order to help a business make more informed decisions.

It also can drain company resources and burden its current staff with routine tasks intended to feed the warehouse machine. Some other disadvantages include the following:

- It takes considerable time and effort to create and maintain the warehouse.
- Gaps in information, caused by human error, can take years to surface, damaging the integrity and usefulness of the information.
- When multiple sources are used, inconsistencies between them can cause information losses.

Advantages

- Provides fact-based analysis on past company performance to inform decision-making.
- Serves as a historical archive of relevant data.
- Can be shared across key departments for maximum usefulness.

Disadvantages

- Creating and maintaining the warehouse is resource-heavy.
- Input errors can damage the integrity of the information archived.
- Use of multiple sources can cause inconsistencies in the data.

What Is a Data Warehouse and What Is It Used for?

A data warehouse is an information storage system for historical data that can be analyzed in numerous ways. Companies and other organizations draw on the data warehouse to gain insight into past performance and plan improvements to their operations.

What Is a Data Warehouse Example?

Consider a company that makes exercise equipment. Its best seller is a stationary bicycle, and it is considering expanding its line and launching a new marketing campaign to support it.

It goes to its data warehouse to understand its current customer better. It can find out whether its customers are predominantly women over 50 or men under 35. It can learn more about the retailers that have been most successful in selling their bikes, and where they're located. It might be able to access in-house survey results and find out what their past customers have liked and disliked about their products.

All of this information helps the company to decide what kind of new model bicycles they want to build and how they will market and advertise them. It's hard information rather than seat-of-the-pants decision-making.

What Are the Stages of Creating a Data Warehouse?

There are at least seven stages to the creation of a data warehouse, according to ITPro Today, an industry publication. They include:

- Determining the business objectives and its key performance indicators.
- Collecting and analyzing the appropriate information.
- Identifying the core business processes that contribute the key data.

- Constructing a conceptual data model that shows how the data are displayed to the end-user.
- Locating the sources of the data and establishing a process for feeding data into the warehouse.
- Establish a tracking duration. Data warehouses can become unwieldy. Many are built with levels of archiving, so that older information is retained in less detail.
- Implementing the plan.

Is SQL a Data Warehouse?

SQL, or Structured Query Language, is a computer language that is used to interact with a database in terms that it can understand and respond to. It contains a number of commands such as "select," "insert," and "update." It is the standard language for relational database management systems.6

A database is not the same as a data warehouse, although both are stores of information. A database is an organized collection of information. A data warehouse is an information archive that is continuously built from multiple sources.7

What Is ETL in a Data Warehouse?

"ETL" stands for "extract, transform, and load." ETL is a data process that combines data from multiple sources into one single data storage unit, which is then loaded into a data warehouse or similar data system. It is used in data analytics and machine learning.

Design guidelines for data warehouse implementation:

Data Warehouse Design

A data warehouse is a single data repository where a record from multiple data sources is integrated for online business analytical processing (OLAP). This implies a data warehouse needs to meet the requirements from all the business stages within the entire organization. Thus, data warehouse design is a hugely complex, lengthy, and hence error-prone process. Furthermore, business analytical functions change over time, which results in changes in the requirements for the systems. Therefore, data warehouse and OLAP systems are dynamic, and the design process is continuous.

Data warehouse design takes a method different from view materialization in the industries. It sees data warehouses as database systems with particular needs such as answering management related queries. The target of the design becomes how the record from multiple data sources should be extracted, transformed, and loaded (ETL) to be organized in a database as the data warehouse.

There are two approaches

1. "top-down" approach
2. "bottom-up" approach

Top-down Design Approach

In the "Top-Down" design approach, a data warehouse is described as a subject-oriented, time-variant, non-volatile and integrated data repository for the entire enterprise data from different sources are validated, reformatted and saved in a normalized (up to 3NF) database as the data warehouse. The data warehouse stores "atomic" information, the data at the lowest level of granularity, from where dimensional data marts can be built by selecting the data required for specific business subjects or particular departments. An approach is a data-driven approach as the information is gathered and integrated first and then business requirements by subjects for building data marts are formulated. The advantage of this method is which it supports a single integrated data source. Thus data marts built from it will have consistency when they overlap.

Advantages of top-down design

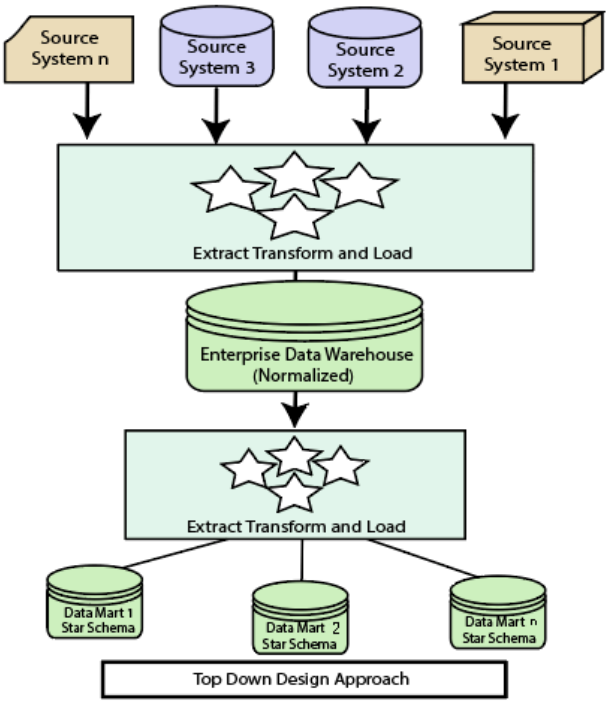
Data Marts are loaded from the data warehouses.

Developing new data mart from the data warehouse is very easy.

Disadvantages of top-down design

This technique is inflexible to changing departmental needs.

The cost of implementing the project is high.



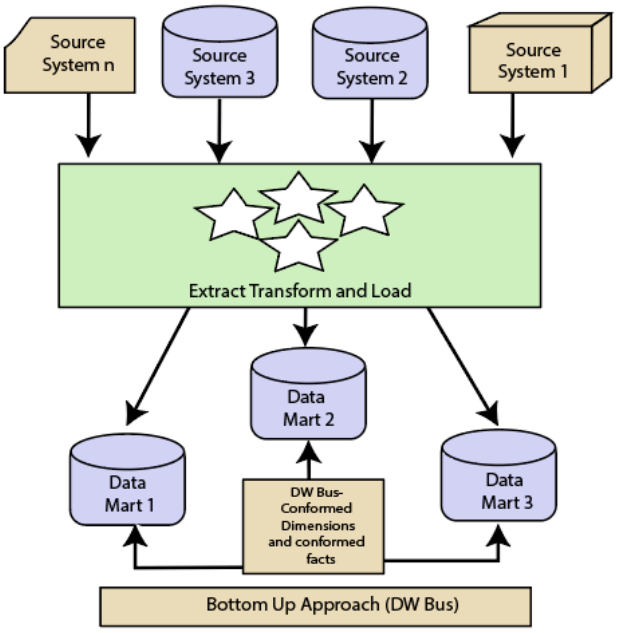
Top Down Design Approach

Bottom-Up Design Approach

In the "Bottom-Up" approach, a data warehouse is described as "a copy of transaction data specifical architecture for query and analysis," term the star schema. In this approach, a data mart is created first to necessary reporting and analytical capabilities for particular business processes (or subjects). Thus it is needed to be a business-driven approach in contrast to Inmon's data-driven approach.

Data marts include the lowest grain data and, if needed, aggregated data too. Instead of a normalized database for the data warehouse, a denormalized dimensional database is adapted to meet the data delivery requirements of data warehouses. Using this method, to use the set of data marts as the enterprise data warehouse, data marts should be built with conformed dimensions in mind, defining that ordinary objects are represented the same in different data marts. The conformed dimensions connected the data marts to form a data warehouse, which is generally called a virtual data warehouse.

The advantage of the "bottom-up" design approach is that it has quick ROI, as developing a data mart, a data warehouse for a single subject, takes far less time and effort than developing an enterprise-wide data warehouse. Also, the risk of failure is even less. This method is inherently incremental. This method allows the project team to learn and grow.



Bottom Up Design Approach

Advantages of bottom-up design

Documents can be generated quickly.

The data warehouse can be extended to accommodate new business units.

It is just developing new data marts and then integrating with other data marts.

Disadvantages of bottom-up design

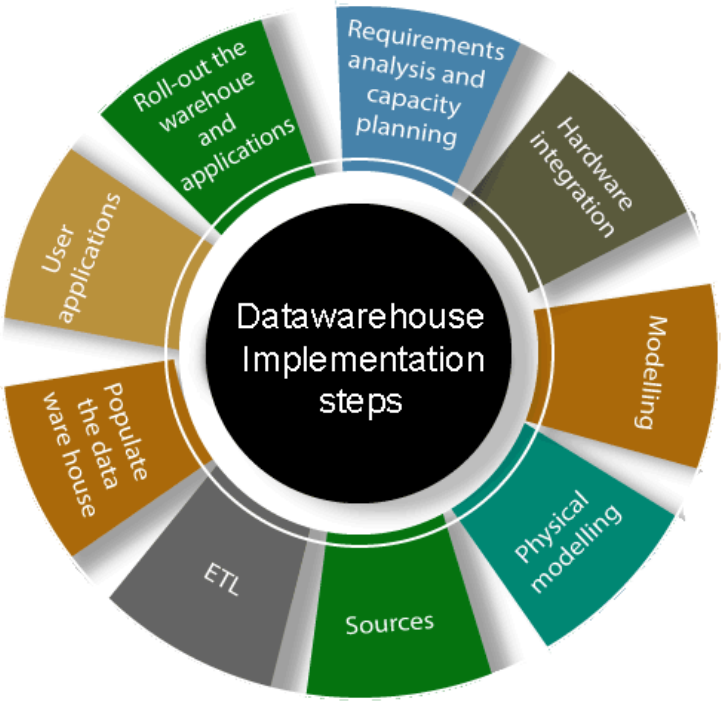
the locations of the data warehouse and the data marts are reversed in the bottom-up approach design.

Differentiate between Top-Down Design Approach and Bottom-Up Design Approach

Top-Down Design Approach	Bottom-Up Design Approach
Breaks the vast problem into smaller subproblems.	Solves the essential low-level problem and integrates them into a higher one.
Inherently architected- not a union of several data marts.	Inherently incremental; can schedule essential data marts first.
Single, central storage of information about the content.	Departmental information stored.
Centralized rules and control.	Departmental rules and control.
It includes redundant information.	Redundancy can be removed.
It may see quick results if implemented with repetitions.	Less risk of failure, favorable return on investment, and proof of techniques.

Data Warehouse Implementation

There are various implementation in data warehouses which are as follows



1. 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. This step will contain be consulting senior management as well as the different stakeholder.

2. 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.

3. Modeling: Modelling 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.

4. 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.

5. 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.

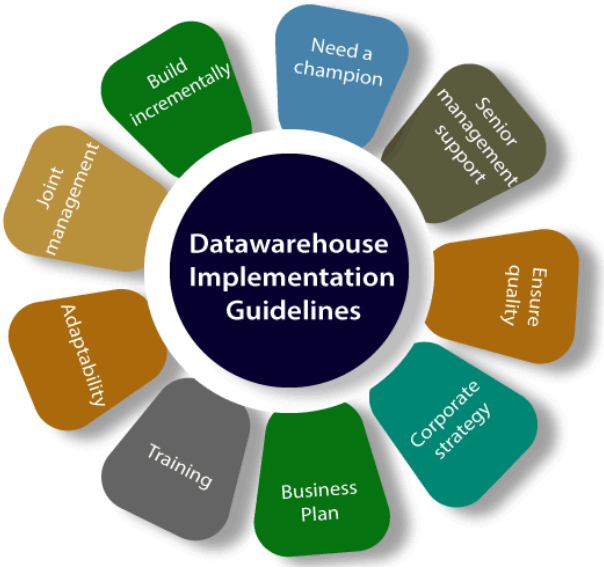
6. ETL: The data from the source system will require to go through an ETL phase. The process of designing and implementing the ETL phase may contain defining a suitable ETL tool vendors and purchasing and implementing the tools. This may contains customize the tool to suit the need of the enterprises.

7. 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.

8. 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.

9. 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.

Implementation Guidelines



1. Build incrementally: Data warehouses must be built incrementally. Generally, it is recommended that a data marts may be created with one particular project in mind, and once it is implemented, several other sections of the enterprise may also want to implement similar systems. An enterprise data warehouses can then be implemented in an iterative manner allowing all data marts to extract information from the data warehouse.

2. Need a champion: A data warehouses project must have a champion who is active to carry out considerable researches into expected price and benefit of the project. Data warehousing projects requires inputs from many units in an enterprise and therefore needs to be driven by someone who is needed for interacting with people in the enterprises and can actively persuade colleagues.

3. Senior management support: A data warehouses project must be fully supported by senior management. Given the resource-intensive feature of such project and the time they can take to implement, a warehouse project signal for a sustained commitment from senior management.

4. Ensure quality: The only record that has been cleaned and is of a quality that is implicit by the organizations should be loaded in the data warehouses.

5. Corporate strategy: A data warehouse project must be suitable for corporate strategies and business goals. The purpose of the project must be defined before the beginning of the projects.

6. Business plan: The financial costs (hardware, software, and peopleware), expected advantage, and a project plan for a data warehouses project must be clearly outlined and understood by all stakeholders. Without such understanding, rumors about expenditure and benefits can become the only sources of data, subversion the projects.

7. Training: Data warehouses projects must not overlook data warehouses training requirements. For a data warehouses project to be successful, the customers must be trained to use the warehouses and to understand its capabilities.

8. Adaptability: The project should build in flexibility so that changes may be made to the data warehouses if and when required. Like any system, a data warehouse will require to change, as the needs of an enterprise change.

9. Joint management: The project must be handled by both IT and business professionals in the enterprise. To ensure that proper communication with the stakeholder and which the project is the target for assisting the enterprise's business, the business professional must be involved in the project along with technical professionals.

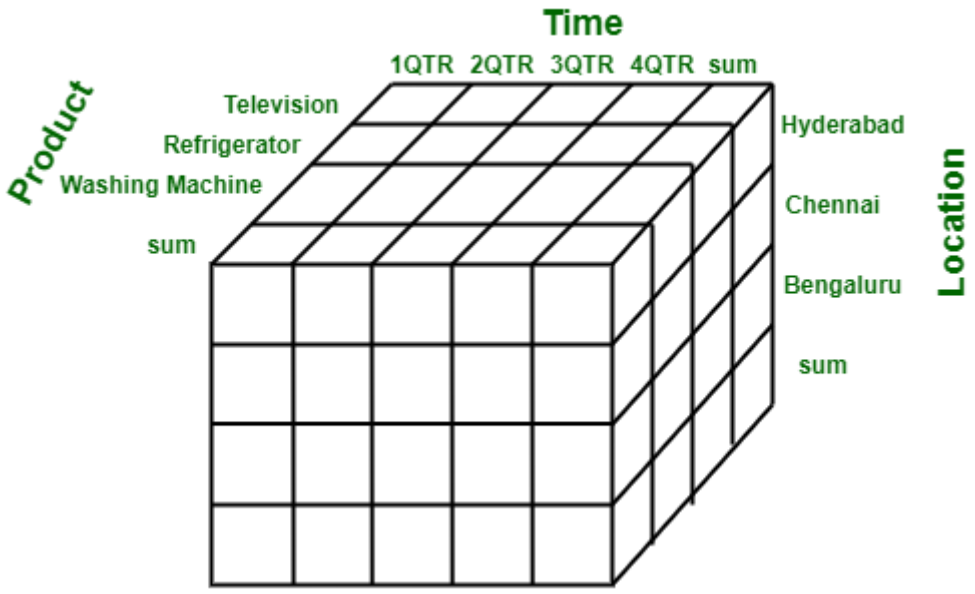
Multidimensional Models

The multi-Dimensional Data Model is a method which is used for ordering data in the database along with good arrangement and assembling of the contents in the database.

The Multi Dimensional Data Model allows customers to interrogate analytical questions associated with market or business trends, unlike relational databases which allow customers to access data in the form of queries. They allow users to rapidly receive answers to the requests which they made by creating and examining the data comparatively fast.

OLAP (online analytical processing) and data warehousing uses multi dimensional databases. It is used to show multiple dimensions of the data to users.

It represents data in the form of data cubes. Data cubes allow to model and view the data from many dimensions and perspectives. It is defined by dimensions and facts and is represented by a fact table. Facts are numerical measures and fact tables contain measures of the related dimensional tables or names of the facts.



Multidimensional Data Representation

Working on a Multidimensional Data Model

On the basis of the pre-decided steps, the Multidimensional Data Model works.

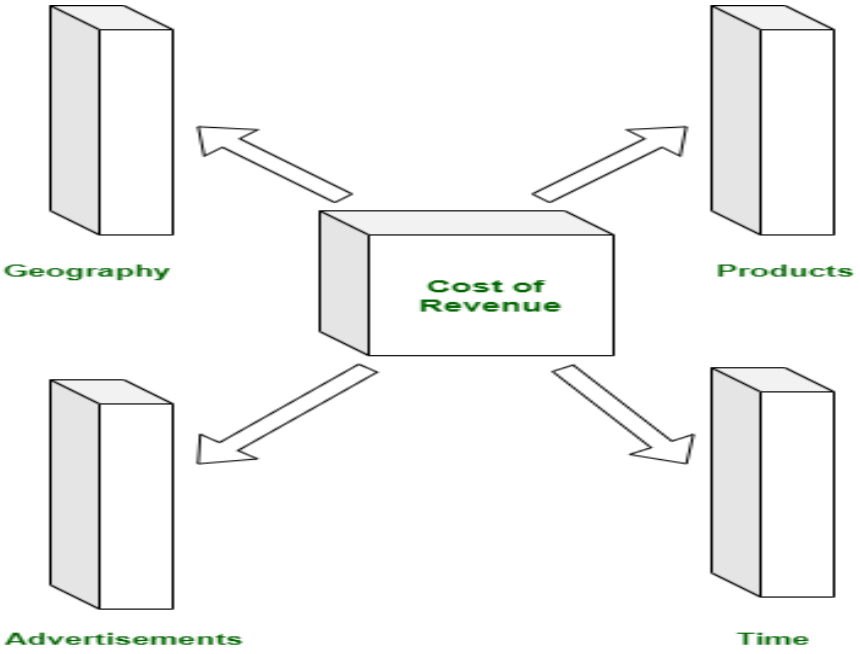
The following stages should be followed by every project for building a Multi-Dimensional Data Model :

Stage 1 : Assembling data from the client : In first stage, a Multi-Dimensional Data Model collects correct data from the client. Mostly, software professionals provide simplicity to the client about the range of data which can be gained with the selected technology and collect the complete data in detail.
Stage 2 : Grouping different segments of the system : In the second stage, the Multi Dimensional Data Model recognizes and classifies all the data to the respective section they belong to and also builds it problem-free to apply step by step.
Stage 3 : Noticing the different proportions : In the third stage, it is the basis on which the design of the system is based. In this stage, the main factors are recognized according to the user's point of view. These factors are also known as "Dimensions".
Stage 4 : Preparing the actual-time factors and their respective qualities : In the fourth stage, the factors which are recognized in the previous step are used further for identifying the related qualities. These qualities are also known as **"attributes"** in the database.
Stage 5 : Finding the actuality of factors which are listed previously and their qualities : In the fifth stage, A Multi Dimensional Data Model separates and differentiates the actuality from the factors which are collected by it. These actually play a significant role in the arrangement of a Multi Dimensional Data Model.

Stage 6 : Building the Schema to place the data, with respect to the information collected from the steps above : In the sixth stage, on the basis of the data which was collected previously, a Schema is built.

For Example :

1. Let us take the example of a firm. The revenue cost of a firm can be recognized on the basis of different factors such as geographical location of firm's workplace, products of the firm, advertisements done, time utilized to flourish a product, etc.



Example 1

2. Let us take the example of the data of a factory which sells products per quarter in Bangalore. The data is represented in the table given below :

Location = "Bangalore"				
Time (quarter)	Type of item			
	Jam	Bread	Sugar	Milk
Q1	350	389	35	50
Q2	260	528	50	90
Q3	483	256	20	60
Q4	436	396	15	40

2D factory data

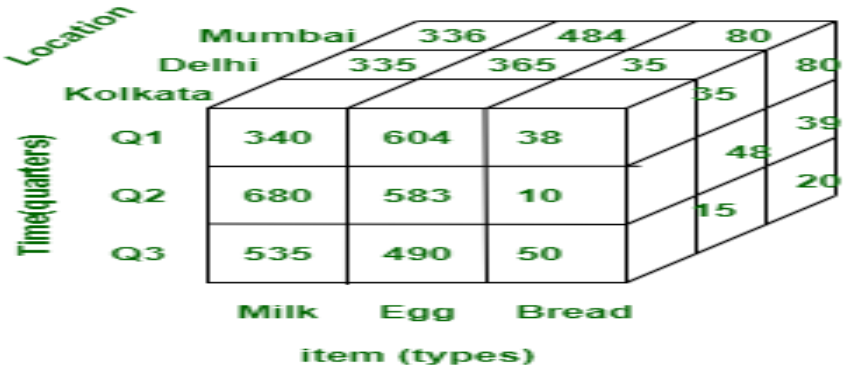
In the above given presentation, the factory's sales for Bangalore are, for the time dimension, which is organized into quarters and the dimension of items, which is sorted according to the kind of item which is sold. The facts here are represented in rupees (in thousands).

Now, if we desire to view the data of the sales in a three-dimensional **table**, then it is represented in the diagram given below. Here the data of the sales is represented as a two **dimensional table**. Let us consider the data according to item, time and location (like Kolkata, Delhi, Mumbai). Here is the table :

	Location="Kolkata"			Location="Delhi"			Location="Mumbai"		
	item			item			item		
Time	Milk	Egg	Bread	Milk	Egg	Bread	Milk	Egg	Bread
Q1	340	604	38	335	365	35	336	484	80
Q2	680	583	10	684	490	48	595	594	39
Q3	535	490	50	389	385	15	366	385	20

3D data representation as 2D

This data can be represented in the form of three dimensions conceptually, which is shown in the image below :



3D data representation

Advantages of Multi Dimensional Data Model

The following are the advantages of a multi-dimensional data model:

- A multi-dimensional data model is easy to handle.
- It is easy to maintain.
- Its performance is better than that of normal databases (e.g. relational databases).
- The representation of data is better than traditional databases. That is because the multi-dimensional databases are multi-viewed and carry different types of factors.
- It is workable on complex systems and applications, contrary to the simple one-dimensional database systems.
- The compatibility in this type of database is an upliftment for projects having lower bandwidth for maintenance staff.

Disadvantages of Multi Dimensional Data Model

The following are the disadvantages of a Multi Dimensional Data Model:

- The multi-dimensional Data Model is slightly complicated in nature and it requires professionals to recognize and examine the data in the database.
- During the work of a Multi-Dimensional Data Model, when the system caches, there is a great effect on the working of the system.
- It is complicated in nature due to which the databases are generally dynamic in design.
- The path to achieving the end product is complicated most of the time.
- As the Multi Dimensional Data Model has complicated systems, databases have a large number of databases due to which the system is very insecure when there is a security break.

OLAP (Online Analytical Processing)?

OLAP stands for **On-Line Analytical Processing**. OLAP is a classification of software technology which authorizes analysts, managers, and executives to gain insight into information through fast, consistent, interactive access in a wide variety of possible views of data that has been transformed from raw information to reflect the real dimensionality of the enterprise as understood by the clients.

OLAP implement the multidimensional analysis of business information and support the capability for complex estimations, trend analysis, and sophisticated data modeling. It is rapidly enhancing the essential foundation for Intelligent Solutions containing Business Performance Management, Planning, Budgeting, Forecasting, Financial Documenting, Analysis, Simulation-Models, Knowledge Discovery, and Data Warehouses Reporting. OLAP enables end-clients to perform ad hoc analysis of record in multiple dimensions, providing the insight and understanding they require for better decision making.

Who uses OLAP and Why?

OLAP applications are used by a variety of the functions of an organization.

Finance and accounting:

- Budgeting
- Activity-based costing
- Financial performance analysis
- And financial modeling

Sales and Marketing

- Sales analysis and forecasting
- Market research analysis
- Promotion analysis
- Customer analysis
- Market and customer segmentation

Production

- Production planning
- Defect analysis

OLAP cubes have two main purposes. The first is to provide business users with a data model more intuitive to them than a tabular model. This model is called a Dimensional Model.

The second purpose is to enable fast query response that is usually difficult to achieve using tabular models.

How OLAP Works?

Fundamentally, OLAP has a very simple concept. It pre-calculates most of the queries that are typically very hard to execute over tabular databases, namely aggregation, joining, and grouping. These queries are calculated during a process that is usually called 'building' or 'processing' of the OLAP cube. This process happens overnight, and by the time end users get to work - data will have been updated.

OLAP Guidelines (Dr.E.F.Codd Rule)

Dr E.F. Codd, the "father" of the relational model, has formulated a list of 12 guidelines and requirements as the basis for selecting OLAP systems:

1) Multidimensional Conceptual View: This is the central features of an OLAP system. By needing a multidimensional view, it is possible to carry out methods like slice and dice.

2) Transparency: Make the technology, underlying information repository, computing operations, and the dissimilar nature of source data totally transparent to users. Such transparency helps to improve the efficiency and productivity of the users.

3) Accessibility: It provides access only to the data that is actually required to perform the particular analysis, present a single, coherent, and consistent view to the clients. The OLAP system must map its own logical schema to the heterogeneous physical data stores and perform any necessary transformations. The OLAP operations should be sitting between data sources (e.g., data warehouses) and an OLAP front-end.

4) Consistent Reporting Performance: To make sure that the users do not feel any significant degradation in documenting performance as the number of dimensions or the size of the database increases. That is, the performance of OLAP should not suffer as the number of dimensions is increased. Users must observe consistent run time, response time, or machine utilization every time a given query is run.

5) Client/Server Architecture: Make the server component of OLAP tools sufficiently intelligent that the various clients to be attached with a minimum of effort and integration programming. The server should be capable of mapping and consolidating data between dissimilar databases.

6) Generic Dimensionality: An OLAP method should treat each dimension as equivalent in both is structure and operational capabilities. Additional operational capabilities may be allowed to selected dimensions, but such additional tasks should be grantable to any dimension.

7) Dynamic Sparse Matrix Handling: To adapt the physical schema to the specific analytical model being created and loaded that optimizes sparse matrix handling. When encountering the sparse matrix, the system must be easy to dynamically assume the distribution of the information and adjust the storage and access to obtain and maintain a consistent level of performance.

8) Multiuser Support: OLAP tools must provide concurrent data access, data integrity, and access security.

9) Unrestricted cross-dimensional Operations: It provides the ability for the methods to identify dimensional order and necessarily functions roll-up and drill-down methods within a dimension or across the dimension.

10) Intuitive Data Manipulation: Data Manipulation fundamental the consolidation direction like as reorientation (pivoting), drill-down and roll-up, and another manipulation to be accomplished naturally and precisely via point-and-click and drag and drop methods on the cells of the scientific model. It avoids the use of a menu or multiple trips to a user interface.

11) Flexible Reporting: It implements efficiency to the business clients to organize columns, rows, and cells in a manner that facilitates simple manipulation, analysis, and synthesis of data.

12) Unlimited Dimensions and Aggregation Levels: The number of data dimensions should be unlimited. Each of these common dimensions must allow a practically unlimited number of customer-defined aggregation levels within any given consolidation path.

Types of OLAP Servers

We have four types of OLAP servers −

- Relational OLAP (ROLAP)
- Multidimensional OLAP (MOLAP)
- Hybrid OLAP (HOLAP)
- Specialized SQL Servers

Relational OLAP

ROLAP servers are placed between relational back-end server and client front-end tools. To store and manage warehouse data, ROLAP uses relational or extended-relational DBMS.

ROLAP includes the following −

- Implementation of aggregation navigation logic.
- Optimization for each DBMS back end.
- Additional tools and services.

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 data set is sparse. Therefore, many MOLAP server use two levels of data storage representation to handle dense and sparse data sets.

Hybrid OLAP

Hybrid OLAP is a combination of both ROLAP and MOLAP. It offers higher scalability of ROLAP and faster computation of MOLAP. HOLAP servers allows to store the large data volumes of detailed information. The aggregations are stored separately in MOLAP store.

Specialized SQL Servers

Specialized SQL servers provide advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

OLAP Operations/ Multidimensional view Efficient processing of OLAP Queries

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

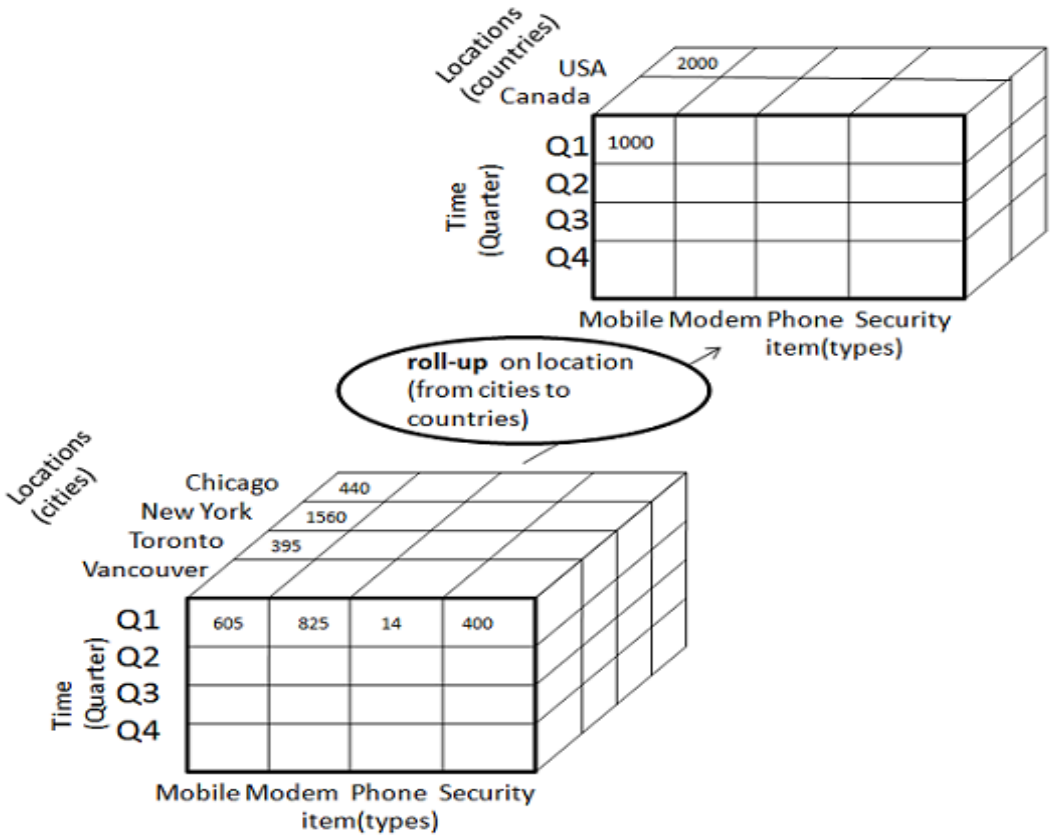
- Drill-down
- Slice and dice
- Pivot (rotate)

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.



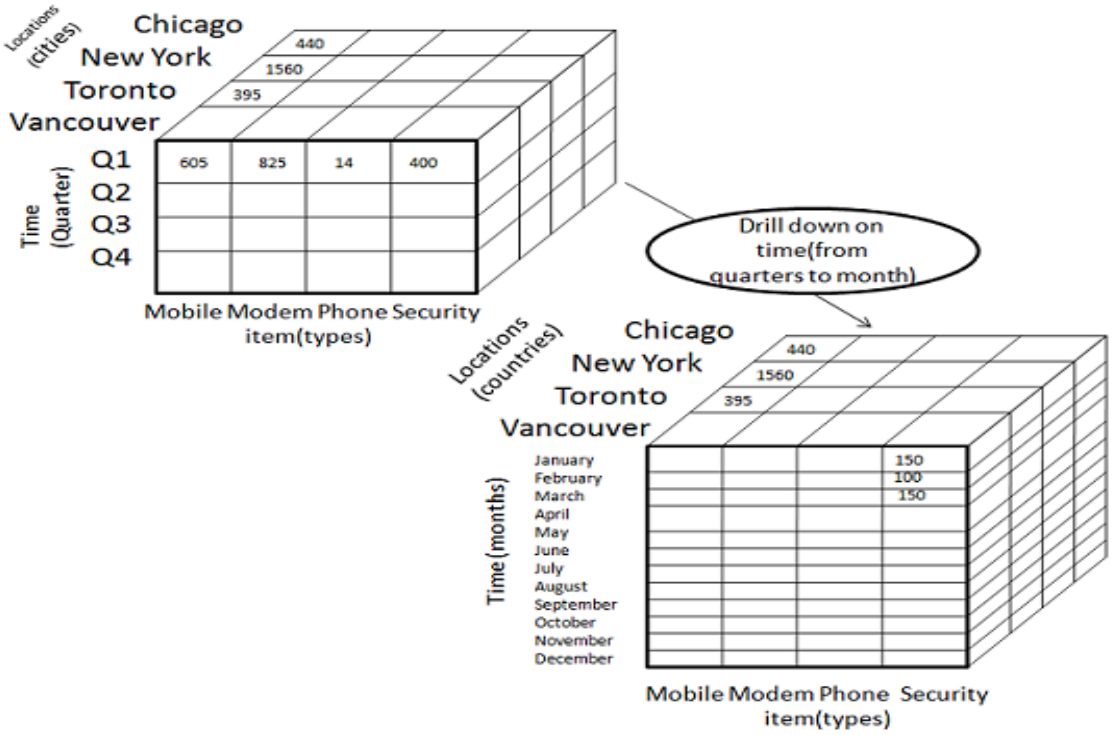
- Roll-up is performed by climbing up a concept hierarchy for the dimension location.
- Initially the concept hierarchy was "street < city < province < country".
- On rolling up, the data is aggregated by ascending the location hierarchy from the level of city to the level of country.
- The data is grouped 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.

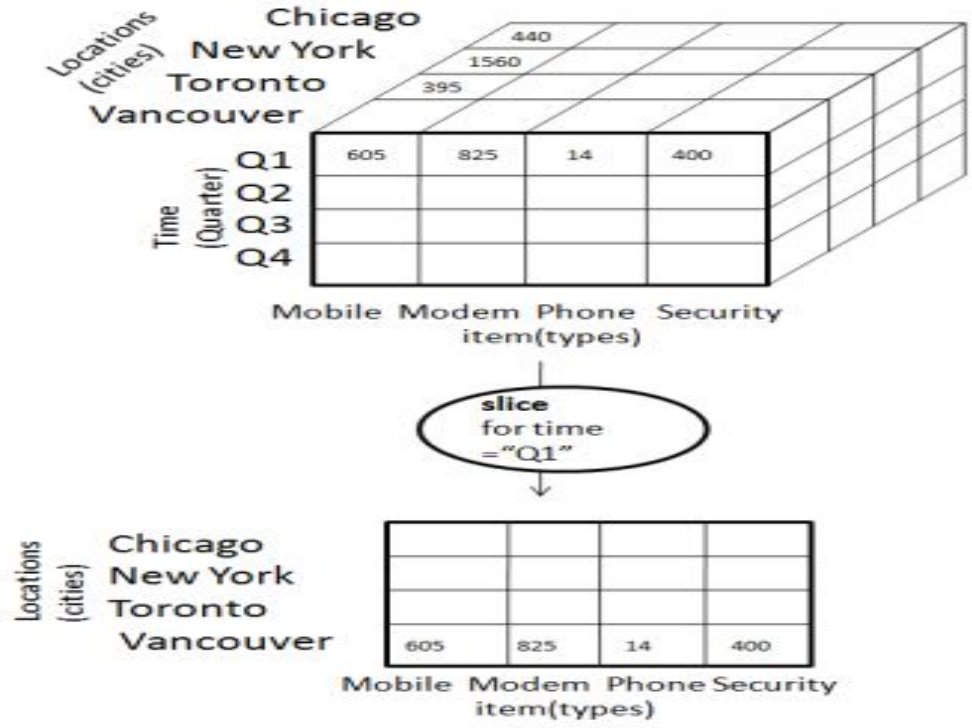
The following diagram illustrates how drill-down works −



- 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

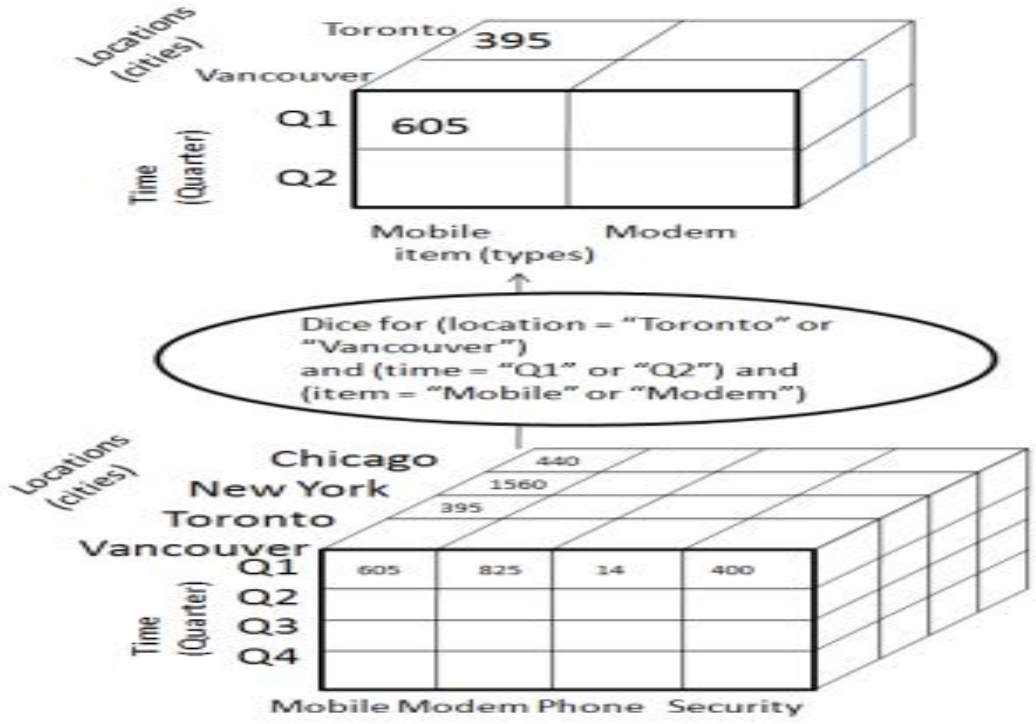
The slice operation selects one particular dimension from a given cube and provides a new sub-cube. Consider the following diagram that shows how slice works.



- 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.

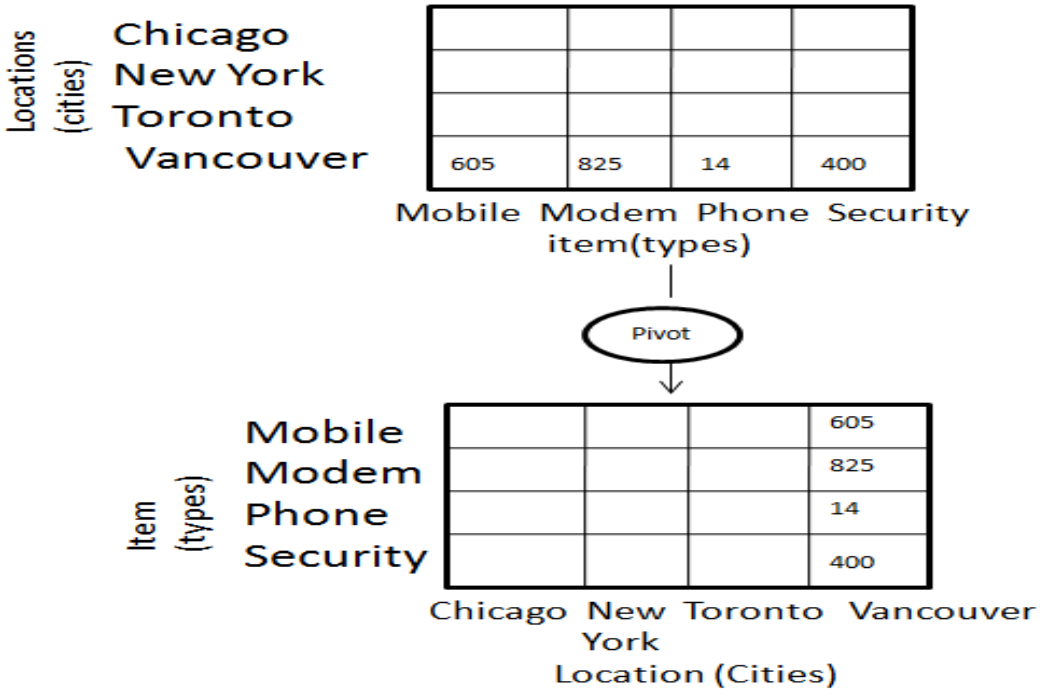


The dice operation on the cube based on the following selection criteria involves three dimensions.

- (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.

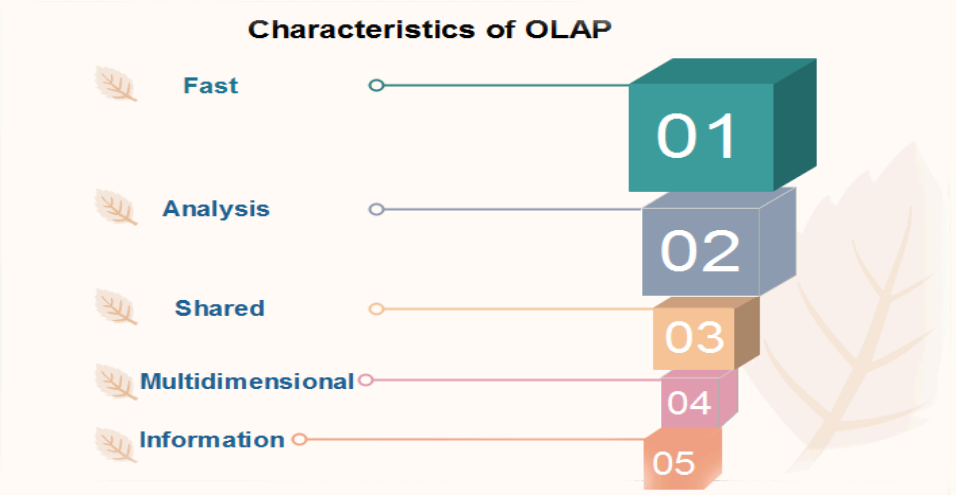


OLAP vs OLTP

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.

Characteristics of OLAP

In the **FASMI characteristics of OLAP methods**, the term derived from the first letters of the characteristics are:



Fast

It defines which the system targeted to deliver the most feedback to the client within about five seconds, with the elementary analysis taking no more than one second and very few taking more than 20 seconds.

Analysis

It defines which the method can cope with any business logic and statistical analysis that is relevant for the function and the user, keep it easy enough for the target client. Although some preprogramming may be needed we do not think it acceptable if all application definitions have to be allow the user to define new Adhoc calculations as part of the analysis and to document on the data in any desired method, without having to program so we excludes products (like Oracle Discoverer) that do not allow the user to define new Adhoc calculation as part of the analysis and to document on the data in any desired product that do not allow adequate end user-oriented calculation flexibility.

Share

It defines which the system tools all the security requirements for understanding and, if multiple write connection is needed, concurrent update location at an appropriated level, not all functions need customer to write data back, but for the increasing number which does, the system should be able to manage multiple updates in a timely, secure manner.

Multidimensional

This is the basic requirement. OLAP system must provide a multidimensional conceptual view of the data, including full support for hierarchies, as this is certainly the most logical method to analyze business and organizations.

Information

The system should be able to hold all the data needed by the applications. Data sparsity should be handled in an efficient manner.

The main characteristics of OLAP are as follows:

1. **Multidimensional conceptual view:** OLAP systems let business users have a dimensional and logical view of the data in the data warehouse. It helps in carrying slice and dice operations.
2. **Multi-User Support:** Since the OLAP techniques are shared, the OLAP operation should provide normal database operations, containing retrieval, update, adequacy control, integrity, and security.
3. **Accessibility:** OLAP acts as a mediator between data warehouses and front-end. The OLAP operations should be sitting between data sources (e.g., data warehouses) and an OLAP front-end.
4. **Storing OLAP results:** OLAP results are kept separate from data sources.
5. **Uniform documenting performance:** Increasing the number of dimensions or database size should not significantly degrade the reporting performance of the OLAP system.
6. OLAP provides for distinguishing between zero values and missing values so that aggregates are computed correctly.
7. OLAP system should ignore all missing values and compute correct aggregate values.
8. OLAP facilitate interactive query and complex analysis for the users.
9. OLAP allows users to drill down for greater details or roll up for aggregations of metrics along a single business dimension or across multiple dimension.
10. OLAP provides the ability to perform intricate calculations and comparisons.
11. OLAP presents results in a number of meaningful ways, including charts and graphs.

Benefits of OLAP: OLAP holds several benefits for businesses: -

1. OLAP helps managers in decision-making through the multidimensional record views that it is efficient in providing, thus increasing their productivity.
2. OLAP functions are self-sufficient owing to the inherent flexibility support to the organized databases.
3. It facilitates simulation of business models and problems, through extensive management of analysis-capabilities.
4. In conjunction with data warehouse, OLAP can be used to support a reduction in the application backlog, faster data retrieval, and reduction in query drag.

Motivations for using OLAP

1) Understanding and improving sales: For enterprises that have much products and benefit a number of channels for selling the product, OLAP can help in finding the most suitable products and the most famous channels. In some methods, it may be feasible to find the most profitable users. **For example,** considering the telecommunication industry and considering only one product, communication minutes, there is a high amount of record if a company want to analyze the sales of products for every hour of the day (24 hours), difference between weekdays and weekends (2 values) and split regions to which calls are made into 50 region.

2) Understanding and decreasing costs of doing business: Improving sales is one method of improving a business, the other method is to analyze cost and to control them as much as suitable without affecting sales. OLAP can assist in analyzing the costs related to sales. In some methods, it may also be feasible to identify expenditures which produce a high return on investments (ROI). **For example,** recruiting a top salesperson may contain high costs, but the revenue generated by the salesperson may justify the investment.

OLAP Servers

Online Analytical Processing (OLAP) refers to a set of software tools used for data analysis in order to make business decisions. OLAP provides a platform for gaining insights from databases retrieved from multiple database systems at the same time. It is based on a multidimensional data model, which enables users to

extract and view data from various perspectives. A multidimensional database is used to store OLAP data. Many Business Intelligence (BI) applications rely on OLAP technology.

Type of OLAP servers:
The three major types of OLAP servers are as follows:

- **ROLAP**
- **MOLAP**
- **HOLAP**

1) **Relational OLAP (ROLAP):**

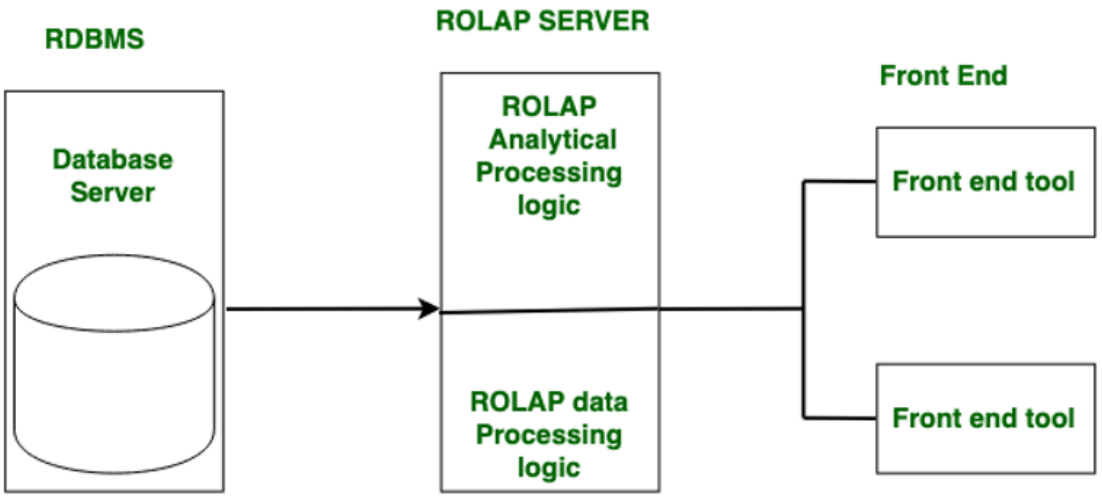
Relational On-Line Analytical Processing (ROLAP) is primarily used for data stored in a relational database, where both the base data and dimension tables are stored as relational tables. ROLAP servers are used to bridge the gap between the relational back-end server and the client's front-end tools. ROLAP servers store and manage warehouse data using RDBMS, and OLAP middleware fills in the gaps.

Benefits:

- It is compatible with data warehouses and OLTP systems.
- The data size limitation of ROLAP technology is determined by the underlying RDBMS. As a result, ROLAP does not limit the amount of data that can be stored.

Limitations:

- SQL functionality is constrained.
- It's difficult to keep aggregate tables up to date.



2) **Multidimensional OLAP (MOLAP):**

Through array-based multidimensional storage engines, Multidimensional On-Line Analytical Processing (MOLAP) supports multidimensional views of data. Storage utilization in multidimensional data stores may be low if the data set is sparse.

MOLAP stores data on discs in the form of a specialized multidimensional array structure. It is used for OLAP, which is based on the arrays' random access capability. Dimension instances determine array elements, and the data or measured value associated with each cell is typically stored in the corresponding array element. The multidimensional array is typically stored in MOLAP in a linear allocation based on nested traversal of the axes in some predetermined order.

However, unlike ROLAP, which stores only records with non-zero facts, all array elements are defined in MOLAP, and as a result, the arrays tend to be sparse, with empty elements occupying a larger portion of them. MOLAP systems typically include provisions such as advanced indexing and hashing to locate data while performing queries for handling sparse arrays, because both storage and retrieval costs are important when evaluating online performance. MOLAP cubes are ideal for slicing and dicing data and can perform complex calculations. When the cube is created, all calculations are pre-generated.

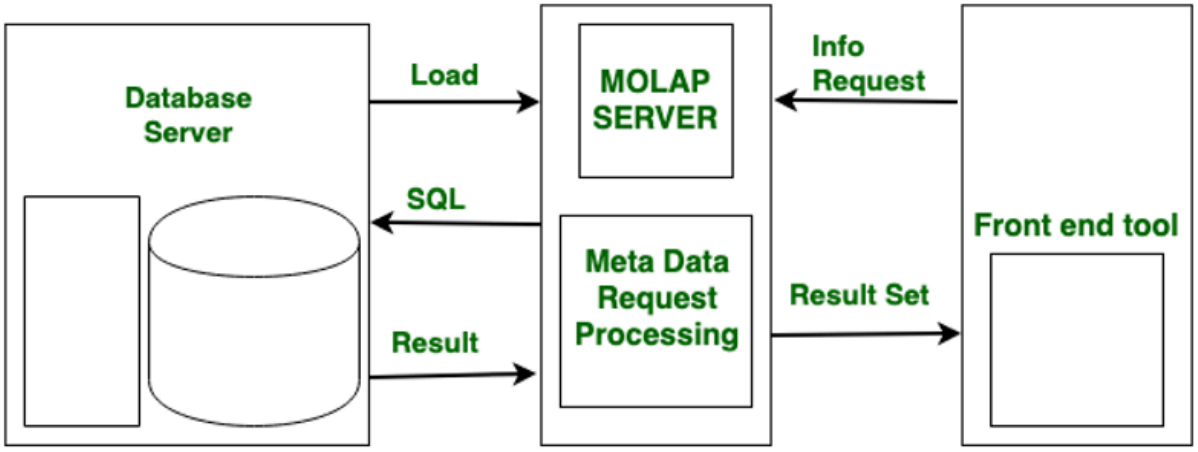
Benefits:

- Suitable for slicing and dicing operations.
- Outperforms ROLAP when data is dense.

- Capable of performing complex calculations.

Limitations:

- It is difficult to change the dimensions without re-aggregating.
- Since all calculations are performed when the cube is built, a large amount of data cannot be stored in the cube itself.



3) **Hybrid OLAP (HOLAP):**

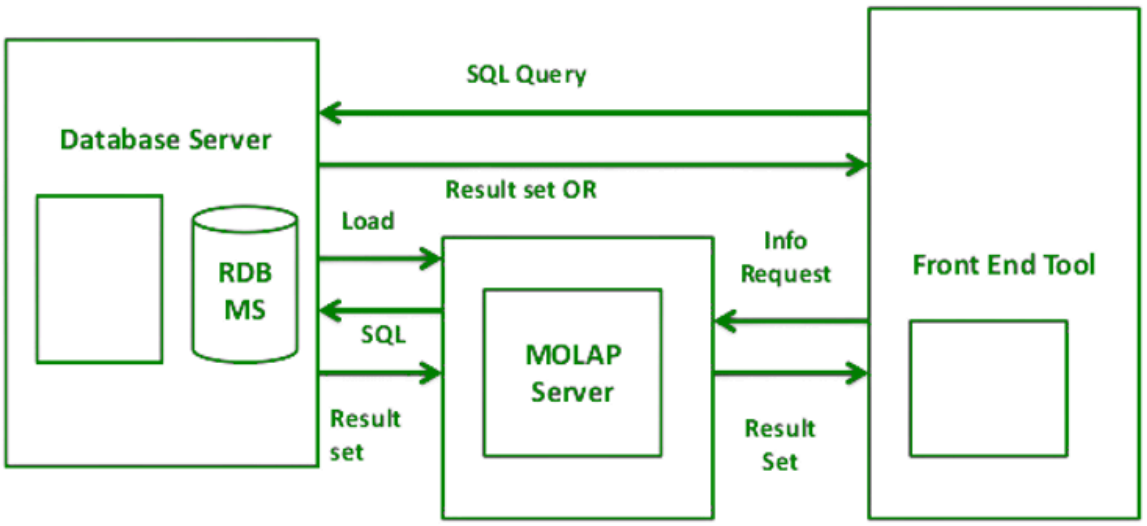
ROLAP and MOLAP are combined in Hybrid On-Line Analytical Processing (HOLAP). HOLAP offers greater scalability than ROLAP and faster computation than MOLAP.HOLAP is a hybrid of ROLAP and MOLAP. HOLAP servers are capable of storing large amounts of detailed data. On the one hand, HOLAP benefits from ROLAP's greater scalability. HOLAP, on the other hand, makes use of cube technology for faster performance and summary-type information. Because detailed data is stored in a relational database, cubes are smaller than MOLAP.

Benefits:

- HOLAP combines the benefits of MOLAP and ROLAP.
- Provide quick access at all aggregation levels.

Limitations

- Because it supports both MOLAP and ROLAP servers, HOLAP architecture is extremely complex.
- There is a greater likelihood of overlap, particularly in their functionalities.



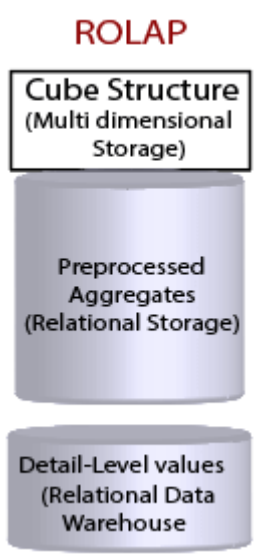
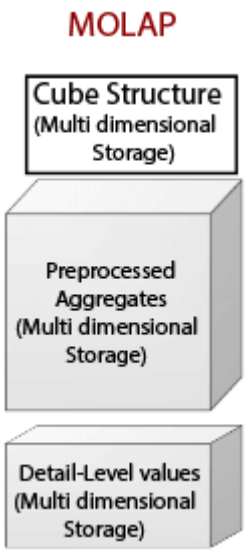
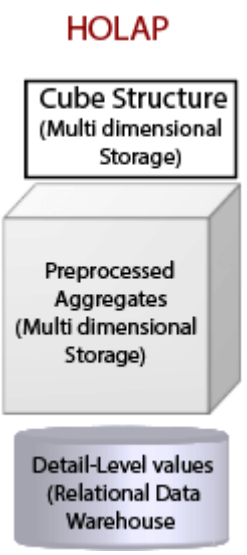
Other types of OLAP include:

- **Web OLAP (WOLAP):** WOLAP refers to an OLAP application that can be accessed through a web browser. WOLAP, in contrast to traditional client/server OLAP applications, is thought to have a three-tiered architecture consisting of three components: a client, middleware, and a database server.

- **Desktop OLAP (DOLAP):** DOLAP is an abbreviation for desktop analytical processing. In that case, the user can download the data from the source and work with it on their desktop or laptop. In comparison to other OLAP applications, functionality is limited. It is less expensive.
- **Mobile OLAP (MOLAP):** Wireless functionality or mobile devices are examples of MOLAP. The user is working and accessing data via mobile devices.
- **Spatial OLAP (SOLAP):** SOLAP egress combines the capabilities of Geographic Information Systems (GIS) and OLAP into a single user interface. SOLAP is created because the data can be alphanumeric, image, or vector. This allows for the quick and easy exploration of data stored in a spatial database.

Difference between ROLAP, MOLAP, and HOLAP(Details)

ROLAP	MOLAP	HOLAP
ROLAP stands for Relational Online Analytical Processing.	MOLAP stands for Multidimensional Online Analytical Processing.	HOLAP stands for Hybrid Online Analytical Processing.
The ROLAP storage mode causes the aggregation of the division to be stored in indexed views in the relational database that was specified in the partition's data source.	The MOLAP storage mode principle the aggregations of the division and a copy of its source information to be saved in a multidimensional operation in analysis services when the separation is processed.	The HOLAP storage mode connects attributes of both MOLAP and ROLAP. Like MOLAP, HOLAP causes the aggregation of the division to be stored in a multidimensional operation in an SQL Server analysis services instance.
ROLAP does not because a copy of the source information to be stored in the Analysis services data folders. Instead, when the outcome cannot be derived from the query cache, the indexed views in the record source are accessed to answer queries.	This MOLAP operation is highly optimize to maximize query performance. The storage area can be on the computer where the partition is described or on another computer running Analysis services. Because a copy of the source information resides in the multidimensional operation, queries can be resolved without accessing the partition's source record.	HOLAP does not causes a copy of the source information to be stored. For queries that access the only summary record in the aggregations of a division, HOLAP is the equivalent of MOLAP.
Query response is frequently slower with ROLAP storage than with the MOLAP or HOLAP storage mode. Processing time is also frequently slower with ROLAP.	Query response times can be reduced substantially by using aggregations. The record in the partition's MOLAP operation is only as current as of the most recent processing of the separation.	Queries that access source record for example, if we want to drill down to an atomic cube cell for which there is no aggregation information must retrieve data from the relational database and will not be as fast as they would be if the

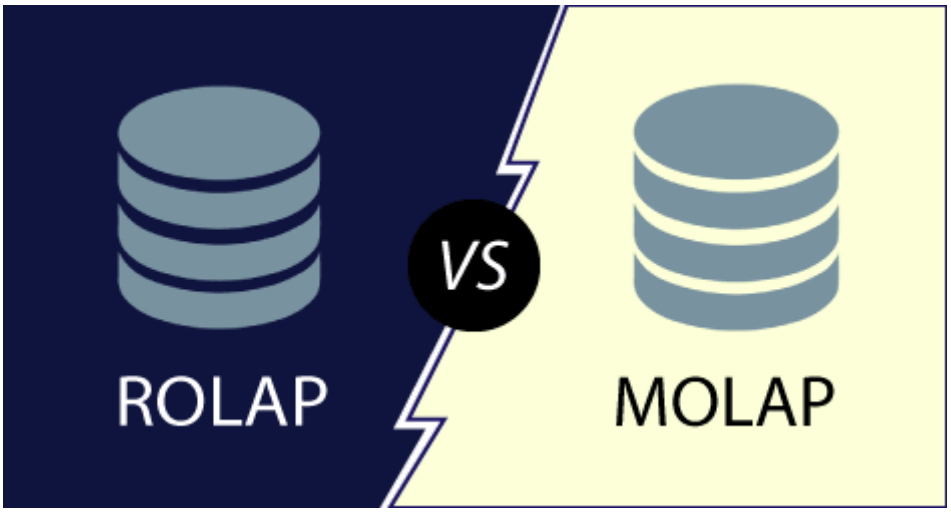
		source information were stored in the MOLAP architecture.
		

Comparison Table of ROLAP vs MOLAP vs HOLAP (Simple Difference)

Basics for comparison	ROLAP	MOLAP	HOLAP
Acronym	Relational online analytical processing	Multi-dimensional online analytical processing	Hybrid online analytical processing
Storage methods	Data is stored on the main data warehouse	Data is stored on the registered database MDDB	Data is stored on the relational databases
Fetching methods	Data is fetched from the main repository	Data is fetched from the Proprietary database	Data is fetched from the relational databases
Data Arrangement	Data is arranged and saved in the form of tables with rows and columns	Data is arranged and stored in the form of data cubes	Data is arranged in multi-dimensional form
Volume	Enormous data is processed	Limited data which is kept in proprietary is processed	Large data can be processed
Technique	It works with SQL	It works with Sparse Matrix technology	It uses both Sparse matrix technology and SQL

Designed view	It has dynamic access	It has a static access	It has dynamic access
Response time	It has Maximum response time	It has Minimum response time	It takes Minimum response time

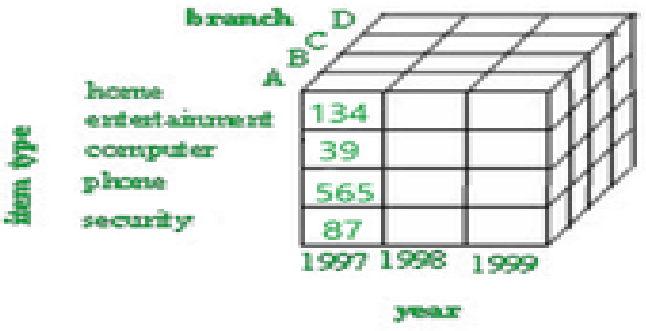
Difference between ROLAP and MOLAP



ROLAP	MOLAP
ROLAP stands for Relational Online Analytical Processing.	MOLAP stands for Multidimensional Online Analytical Processing.
It usually used when data warehouse contains relational data.	It used when data warehouse contains relational as well as non-relational data.
It contains Analytical server.	It contains the MDDB server.
It creates a multidimensional view of data dynamically.	It contains prefabricated data cubes.
It is very easy to implement	It is difficult to implement.
It has a high response time	It has less response time due to prefabricated cubes.
It requires less amount of memory.	It requires a large amount of memory.

Data Cube or OLAP approach in Data Mining

Grouping of data in a multidimensional matrix is called data cubes. In Dataware housing, we generally deal with various multidimensional data models as the data will be represented by multiple dimensions and multiple attributes. This multidimensional data is represented in the data cube as the cube represents a high-dimensional space. The Data cube pictorially shows how different attributes of data are arranged in the data model. Below is the diagram of a general data cube.



The example above is a 3D cube having attributes like branch(A,B,C,D),

item type(home, entertainment, computer, phone, security),

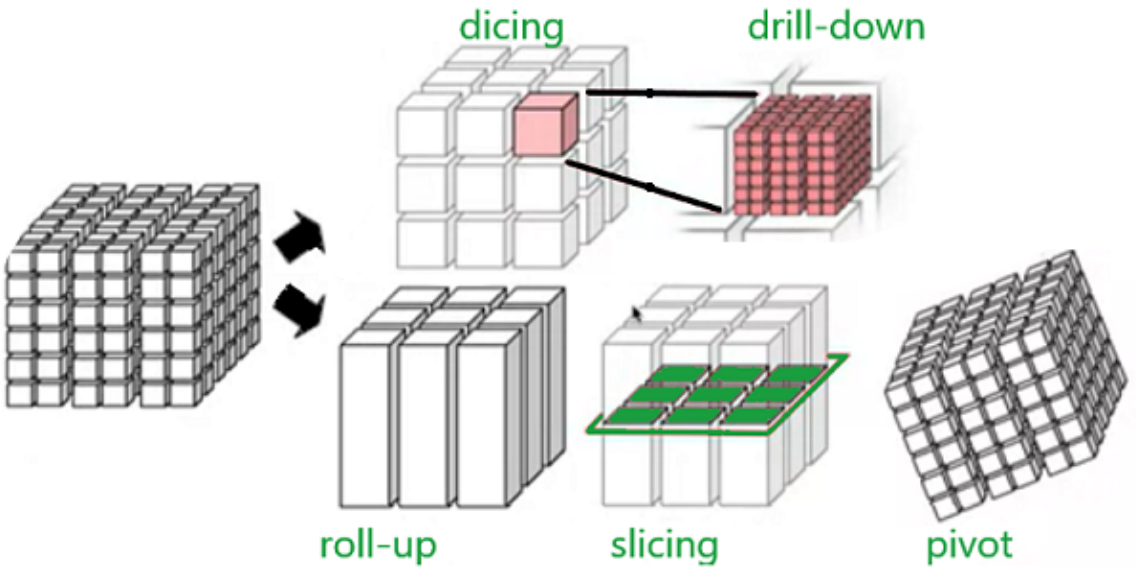
year(1997,1998,1999) .

Data cube classification:

The data cube can be classified into two categories:

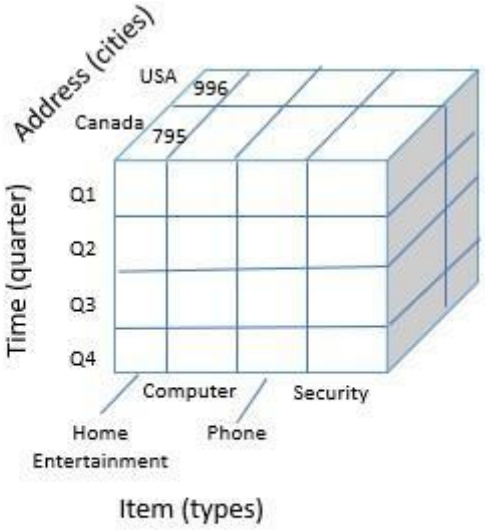
- **Multidimensional data cube:** It basically helps in storing large amounts of data by making use of a multi-dimensional array. It increases its efficiency by keeping an index of each dimension. Thus, dimensional is able to retrieve data fast.
- **Relational data cube:** It basically helps in storing large amounts of data by making use of relational tables. Each relational table displays the dimensions of the data cube. It is slower compared to a Multidimensional Data Cube.

Data cube operations:



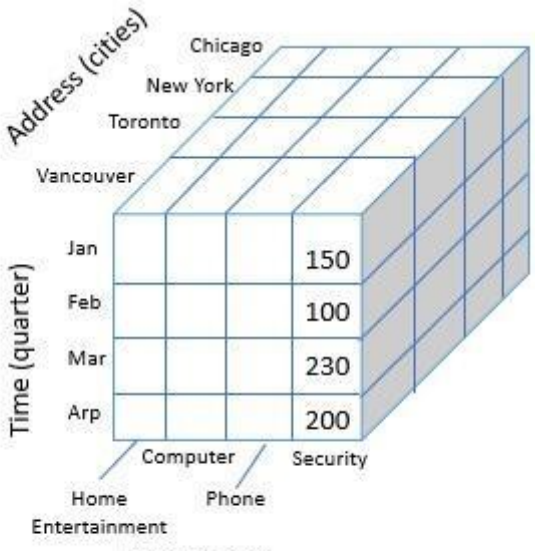
Data cube operations are used to manipulate data to meet the needs of users. These operations help to select particular data for the analysis purpose. There are mainly 5 operations listed below-

- **Roll-up**: operation and aggregate certain similar data attributes having the same dimension together. For example, if the data cube displays the daily income of a customer, we can use a roll-up operation to find the monthly income of his salary.



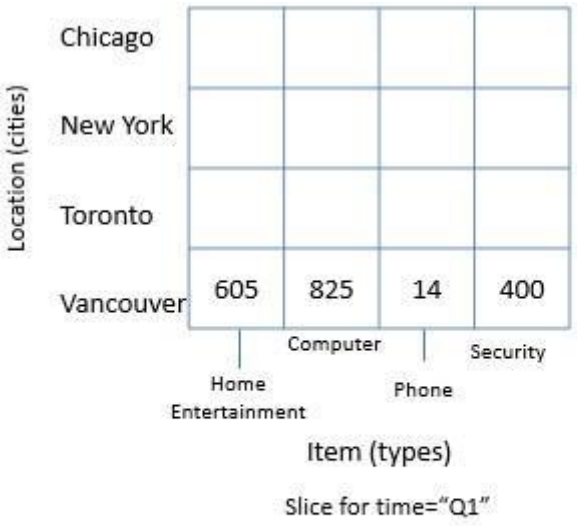
Data Cube Roll-up Operation on Address

- **Drill-down**: this operation is the reverse of the roll-up operation. It allows us to take particular information and then subdivide it further for coarser granularity analysis. It zooms into more detail. For example- if India is an attribute of a country column and we wish to see villages in India, then the drill-down operation splits India into states, districts, towns, cities, villages and then displays the required information.



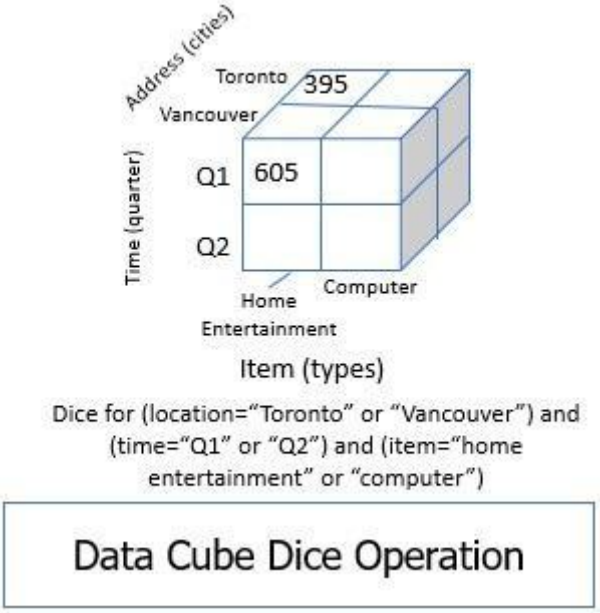
Data Cube Drill-down Operation on Time

- **Slicing**: this operation filters the unnecessary portions. Suppose in a particular dimension, the user doesn't need everything for analysis, rather a particular attribute. For example, country="jamaica", this will display only about jamaica and only display other countries present on the country list.



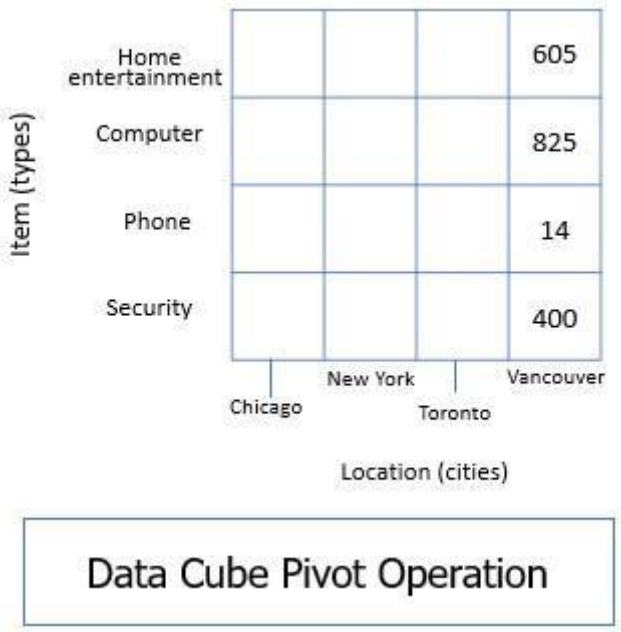
Data Cube Dice Operation

- **Dicing**: this operation does a multidimensional cutting, that not only cuts only one dimension but also can go to another dimension and cut a certain range of it. As a result, it looks more like a subcube out of the whole cube(as depicted in the figure). For example- the user wants to see the annual salary of Jharkhand state employees.



Data Cube Dice Operation

- **Pivot**: this operation is very important from a viewing point of view. It basically transforms the data cube in terms of view. It doesn't change the data present in the data cube. For example, if the user is comparing year versus branch, using the pivot operation, the user can change the viewpoint and now compare branch versus item type.



Data Cube Pivot Operation

Advantages of data cubes:

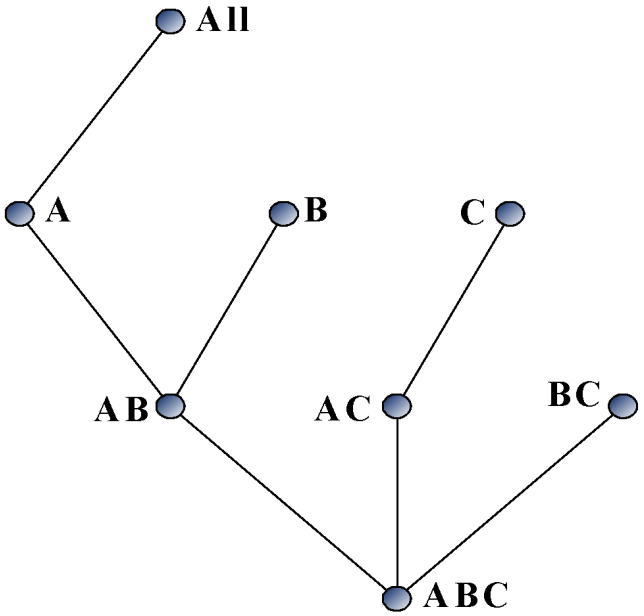
- Helps in giving a summarised view of data.
- Data cubes store large data in a simple way.
- Data cube operation provides quick and better analysis,
- Improve performance of data.

Data Cube Computation Methods

- Multi-Way Array Aggregation.
- BUC.
- Star-Cubing.
- High-Dimensional OLAP.

1) Multi-Way Array Aggregation:

- Array-based "bottom-up" algorithm
- Using multi-dimensional chunks
- No direct tuple comparisons
- Simultaneous aggregation on multiple dimensions
- Intermediate aggregate values are re-used for computing ancestor cuboids
- Cannot do *Apriori* pruning: No iceberg optimization



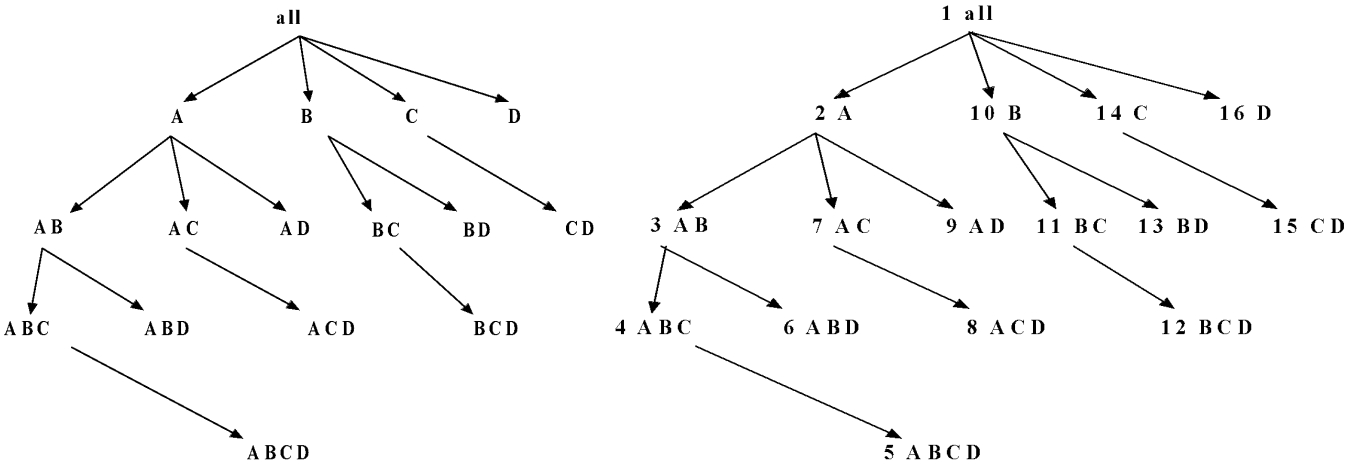
2) Bottom-Up Computation (BUC):

- BUC (Beyer & Ramakrishnan, SIGMOD'99)
- Bottom-up cube computation

(Note: top-down in our view!)

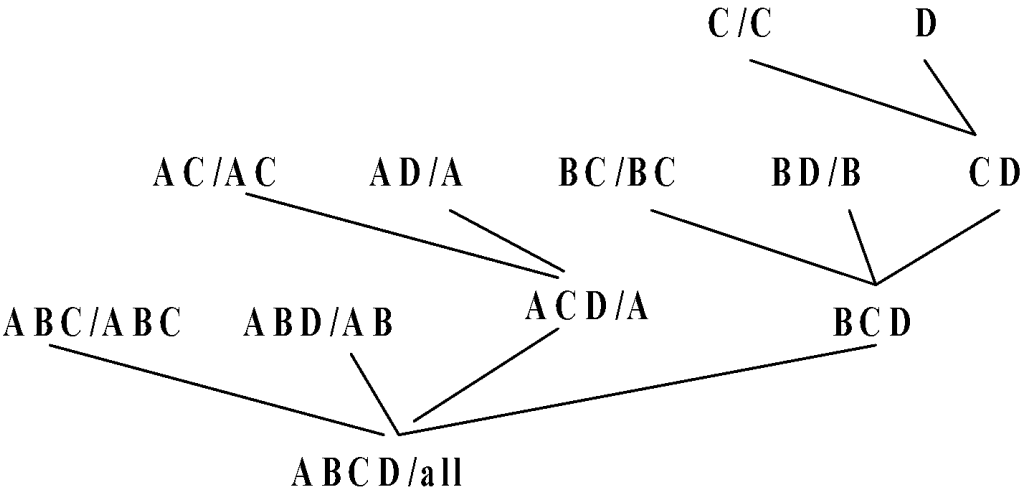
- Divides dimensions into partitions and facilitates iceberg pruning
  - If a partition does not satisfy *min_sup*, its descendants can be pruned
  - If *minsup* = 1 Þ compute full CUBE!

No simultaneous aggregation



3) Star-Cubing: An Integrating Method:

- D. Xin, J. Han, X. Li, B. W. Wah, Star-Cubing: Computing Iceberg Cubes by Top-Down and Bottom-Up Integration, VLDB'03
- ***Explore shared dimensions***
  - E.g., dimension A is the shared dimension of ACD and AD
  - ABD/AB means cuboid ABD has shared dimensions AB
- Allows for shared computations
  - e.g., cuboid AB is computed simultaneously as ABD
- Aggregate in a top-down manner but with the bottom-up sub-layer underneath which will allow Apriori pruning
- Shared dimensions grow in bottom-up fashion



4) High-Dimensional OLAP.

- None of the previous cubing method can handle high dimensionality!
- A database of 600k tuples. Each dimension has cardinality of 100 and *zipf* of 2.

- X. Li, J. Han, and H. Gonzalez, High-Dimensional OLAP: A Minimal Cubing Approach, VLDB'04
- Challenge to current cubing methods:
  - The "curse of dimensionality" problem

- Iceberg cube and compressed cubes: only delay the inevitable explosion
- Full materialization: still significant overhead in accessing results on disk

- High-D OLAP is needed in applications
  - Science and engineering analysis
  - Bio-data analysis: thousands of genes

Statistical surveys: hundreds of variables

- Observation: OLAP occurs only on a small subset of dimensions at a time
- Semi-Online Computational Model
  1. Partition the set of dimensions into **shell fragments**
  2. Compute data cubes for each shell fragment while retaining **inverted indices** or **value-list indices**
  3. Given the pre-computed **fragment cubes**, dynamically compute cube cells of the high-dimensional data cube *online*

Data mining:

Data Mining Tutorial

The data mining tutorial provides basic and advanced concepts of data mining. Our data mining tutorial is designed for learners and experts.

Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called ***Knowledge Discovery in Database (KDD)***. The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation.

Our Data mining tutorial includes all topics of Data mining such as applications, Data mining vs Machine learning, Data mining tools, Social Media Data mining, Data mining techniques, Clustering in data mining, Challenges in Data mining, etc.

What is Data Mining?

The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.

In other words, we can say that Data Mining is the process of investigating hidden patterns of information to various perspectives for categorization into useful data, which is collected and assembled in particular areas such as data warehouses, efficient analysis, data mining algorithm, helping decision making and other data requirement to eventually cost-cutting and generating revenue.

Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures. Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events. Data Mining is also called Knowledge Discovery of Data (KDD).

Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information.

Data Mining is similar to Data Science carried out by a person, in a specific situation, on a particular data set, with an objective. This process includes various types of services such as text mining, web mining, audio and video mining, pictorial data mining, and social media mining. It is done through software that is simple or highly specific. By outsourcing data mining, all the work can be done faster with low operation costs. Specialized firms can also use new technologies to collect data that is impossible to locate manually. There are tonnes of information available on various platforms, but very little knowledge is accessible. The biggest challenge is to analyze the data to extract important information that can be used to solve a problem or for company development. There are many powerful instruments and techniques available to mine data and find better insight from it.

Types of Data Mining

Data mining can be performed on the following types of data:

Relational Database:

A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the database tables. Tables convey and share information, which facilitates data searchability, reporting, and organization.

Data warehouses:

A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights. The huge amount of data comes from multiple places such as Marketing and Finance. The extracted data is utilized for analytical purposes and helps in decision- making for a business organization. The data warehouse is designed for the analysis of data rather than transaction processing.

Data Repositories:

The Data Repository generally refers to a destination for data storage. However, many IT professionals utilize the term more clearly to refer to a specific kind of setup within an IT structure. For example, a group of databases, where an organization has kept various kinds of information.

Object-Relational Database:

A combination of an object-oriented database model and relational database model is called an object-relational model. It supports Classes, Objects, Inheritance, etc.

One of the primary objectives of the Object-relational data model is to close the gap between the Relational database and the object-oriented model practices frequently utilized in many programming languages, for example, C++, Java, C#, and so on.

Transactional Database:

A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database activities.

Advantages of Data Mining

- The Data Mining technique enables organizations to obtain knowledge-based data.
- Data mining enables organizations to make lucrative modifications in operation and production.
- Compared with other statistical data applications, data mining is a cost-efficient.
- Data Mining helps the decision-making process of an organization.
- It Facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
- It can be induced in the new system as well as the existing platforms.
- It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

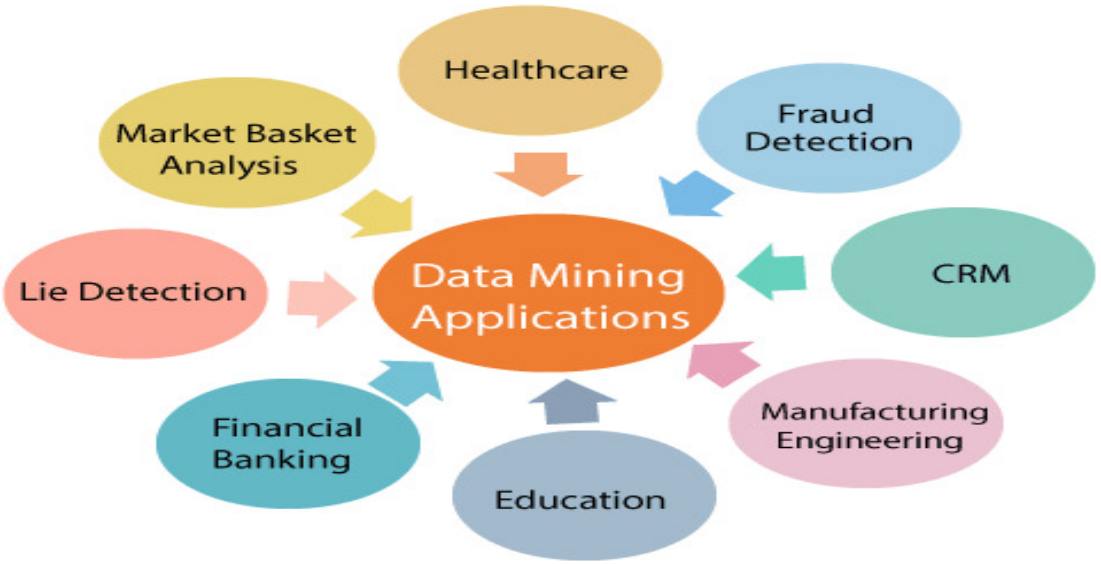
Disadvantages of Data Mining

- There is a probability that the organizations may sell useful data of customers to other organizations for money. As per the report, American Express has sold credit card purchases of their customers to other organizations.
- Many data mining analytics software is difficult to operate and needs advance training to work on.

- Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.
- The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

Data Mining Applications

Data Mining is primarily used by organizations with intense consumer demands- Retail, Communication, Financial, marketing company, determine price, consumer preferences, product positioning, and impact on sales, customer satisfaction, and corporate profits. Data mining enables a retailer to use point-of-sale records of customer purchases to develop products and promotions that help the organization to attract the customer.



These are the following areas where data mining is widely used:

Data Mining in Healthcare:

Data mining in healthcare has excellent potential to improve the health system. It uses data and analytics for better insights and to identify best practices that will enhance health care services and reduce costs. Analysts use data mining approaches such as Machine learning, Multi-dimensional database, Data visualization, Soft computing, and statistics. Data Mining can be used to forecast patients in each category. The procedures ensure that the patients get intensive care at the right place and at the right time. Data mining also enables healthcare insurers to recognize fraud and abuse.

Data Mining in Market Basket Analysis:

Market basket analysis is a modeling method based on a hypothesis. If you buy a specific group of products, then you are more likely to buy another group of products. This technique may enable the retailer to understand the purchase behavior of a buyer. This data may assist the retailer in understanding the requirements of the buyer and altering the store's layout accordingly. Using a different analytical comparison of results between various stores, between customers in different demographic groups can be done.

Data mining in Education:

Education data mining is a newly emerging field, concerned with developing techniques that explore knowledge from the data generated from educational Environments. EDM objectives are recognized as affirming student's future learning behavior, studying the impact of educational support, and promoting learning science. An organization can use data mining to make precise decisions and also to predict the results of the student. With the results, the institution can concentrate on what to teach and how to teach.

Data Mining in Manufacturing Engineering:

Knowledge is the best asset possessed by a manufacturing company. Data mining tools can be beneficial to find patterns in a complex manufacturing process. Data mining can be used in system-level designing to obtain the relationships between product architecture, product portfolio, and data needs of the customers. It can also be used to forecast the product development period, cost, and expectations among the other tasks.

Data Mining in CRM (Customer Relationship Management):

Customer Relationship Management (CRM) is all about obtaining and holding Customers, also enhancing customer loyalty and implementing customer-oriented strategies. To get a decent relationship with the customer, a business organization needs to collect data and analyze the data. With data mining technologies, the collected data can be used for analytics.

Data Mining in Fraud detection:

Billions of dollars are lost to the action of frauds. Traditional methods of fraud detection are a little bit time consuming and sophisticated. Data mining provides meaningful patterns and turning data into information. An ideal fraud detection system should protect the data of all the users. Supervised methods consist of a collection of sample records, and these records are classified as fraudulent or non-fraudulent. A model is constructed using this data, and the technique is made to identify whether the document is fraudulent or not.

Data Mining in Lie Detection:

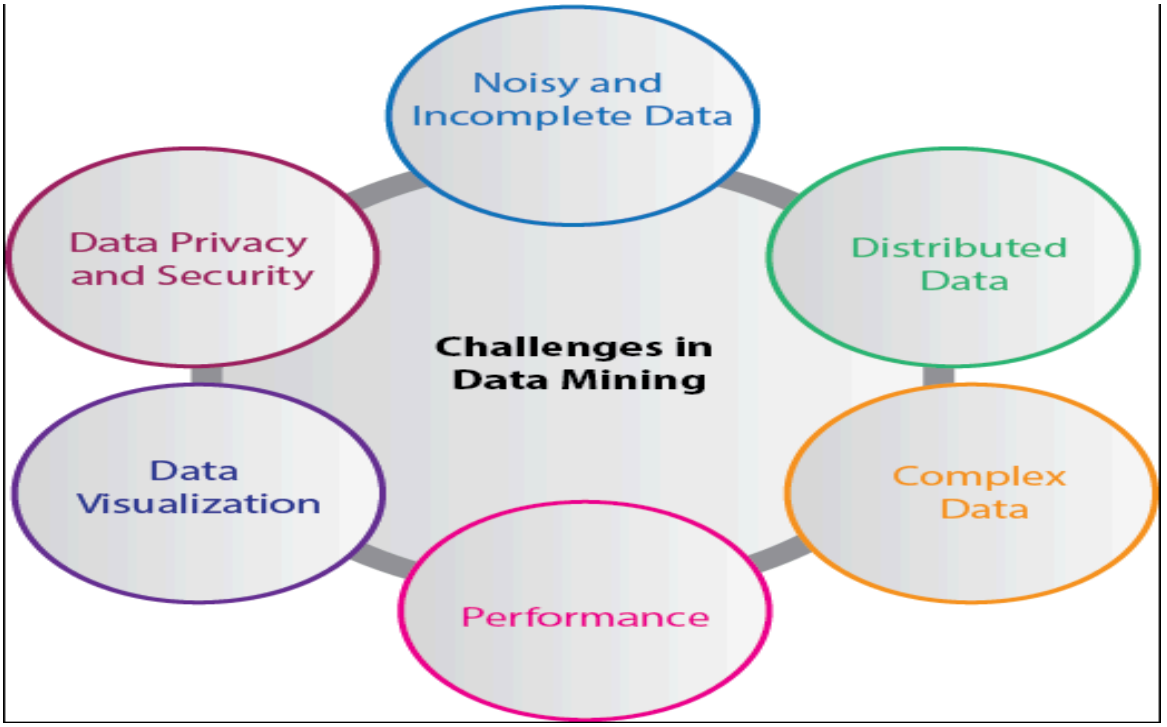
Apprehending a criminal is not a big deal, but bringing out the truth from him is a very challenging task. Law enforcement may use data mining techniques to investigate offenses, monitor suspected terrorist communications, etc. This technique includes text mining also, and it seeks meaningful patterns in data, which is usually unstructured text. The information collected from the previous investigations is compared, and a model for lie detection is constructed.

Data Mining Financial Banking:

The Digitalization of the banking system is supposed to generate an enormous amount of data with every new transaction. The data mining technique can help bankers by solving business-related problems in banking and finance by identifying trends, casualties, and correlations in business information and market costs that are not instantly evident to managers or executives because the data volume is too large or are produced too rapidly on the screen by experts. The manager may find these data for better targeting, acquiring, retaining, segmenting, and maintain a profitable customer.

Challenges of Implementation in Data mining

Although data mining is very powerful, it faces many challenges during its execution. Various challenges could be related to performance, data, methods, and techniques, etc. The process of data mining becomes effective when the challenges or problems are correctly recognized and adequately resolved.



Incomplete and noisy data:

The process of extracting useful data from large volumes of data is data mining. The data in the real-world is heterogeneous, incomplete, and noisy. Data in huge quantities will usually be inaccurate or unreliable. These problems may occur due to data measuring instrument or because of human errors. Suppose a retail chain collects phone numbers of customers who spend more than $ 500, and the accounting employees put the information into their system. The person may make a digit mistake when entering the phone number, which results in incorrect data. Even some customers may not be willing to disclose their phone numbers, which results in incomplete data. The data could get changed due to human or system error. All these consequences (noisy and incomplete data)makes data mining challenging.

Data Distribution:

Real-worlds data is usually stored on various platforms in a distributed computing environment. It might be in a database, individual systems, or even on the internet. Practically, It is a quite tough task to make all the data to a centralized data repository mainly due to organizational and technical concerns. For example, various regional offices may have their servers to store their data. It is not feasible to store, all the data from all the offices on a central server. Therefore, data mining requires the development of tools and algorithms that allow the mining of distributed data.

Complex Data:

Real-world data is heterogeneous, and it could be multimedia data, including audio and video, images, complex data, spatial data, time series, and so on. Managing these various types of data and extracting useful information is a tough task. Most of the time, new technologies, new tools, and methodologies would have to be refined to obtain specific information.

Performance:

The data mining system's performance relies primarily on the efficiency of algorithms and techniques used. If the designed algorithm and techniques are not up to the mark, then the efficiency of the data mining process will be affected adversely.

Data Privacy and Security:

Data mining usually leads to serious issues in terms of data security, governance, and privacy. For example, if a retailer analyzes the details of the purchased items, then it reveals data about buying habits and preferences of the customers without their permission.

Data Visualization:

In data mining, data visualization is a very important process because it is the primary method that shows the output to the user in a presentable way. The extracted data should convey the exact meaning of what it intends to express. But many times, representing the information to the end-user in a precise and easy way is difficult. The input data and the output information being complicated, very efficient, and successful data visualization processes need to be implemented to make it successful.

Tasks and Functionalities of Data Mining

Data mining tasks are designed to be semi-automatic or fully automatic and on large data sets to uncover patterns such as groups or clusters, unusual or over the top data called anomaly detection and dependencies such as association and sequential pattern. Once patterns are uncovered, they can be thought of as a summary of the input data, and further analysis may be carried out using Machine Learning and Predictive analytics. For example, the data mining step might help identify multiple groups in the data that a decision support system can use. Note that data collection, preparation, reporting are not part of data mining.

There is a lot of confusion between data mining and data analysis. Data mining functions are used to define the trends or correlations contained in data mining activities. While data analysis is used to test statistical

models that fit the dataset, for example, analysis of a marketing campaign, data mining uses Machine Learning and mathematical and statistical models to discover patterns hidden in the data. In comparison, data mining activities can be divided into two categories:

- **Descriptive Data Mining:** It includes certain knowledge to understand what is happening within the data without a previous idea. The common data features are highlighted in the data set. For example, count, average etc.
- **Predictive Data Mining:** It helps developers to provide unlabeled definitions of attributes. With previously available or historical data, data mining can be used to make predictions about critical business metrics based on data's linearity. For example, predicting the volume of business next quarter based on performance in the previous quarters over several years or judging from the findings of a patient's medical examinations that is he suffering from any particular disease.

Functionalities of Data Mining

Data mining functionalities are used to represent the type of patterns that have to be discovered in data mining tasks. Data mining tasks can be classified into two types: descriptive and predictive. Descriptive mining tasks define the common features of the data in the database, and the predictive mining tasks act in inference on the current information to develop predictions.

Data mining is extensively used in many areas or sectors. It is used to predict and characterize data. But the ultimate objective in **Data Mining Functionalities** is to observe the various trends in data mining. There are several data mining functionalities that the organized and scientific methods offer, such as:

1. Class/Concept Descriptions

A class or concept implies there is a data set or set of features that define the class or a concept. A class can be a category of items on a shop floor, and a concept could be the abstract idea on which data may be categorized like products to be put on clearance sale and non-sale products. There are two concepts here, one that helps with grouping and the other that helps in differentiating.

- **Data Characterization:** This refers to the summary of general characteristics or features of the class, resulting in specific rules that define a target class. A data analysis technique called Attribute-oriented Induction is employed on the data set for achieving characterization.
- **Data Discrimination:** Discrimination is used to separate distinct data sets based on the disparity in attribute values. It compares features of a class with features of one or more contrasting classes.g., bar charts, curves and pie charts.

2. Mining Frequent Patterns

One of the functions of data mining is finding data patterns. Frequent patterns are things that are discovered to be most common in data. Various types of frequency can be found in the dataset.

- **Frequent item set:**This term refers to a group of items that are commonly found together, such as milk and sugar.
- **Frequent substructure:** It refers to the various types of data structures that can be combined with an item set or subsequences, such as trees and graphs.
- **Frequent Subsequence:** A regular pattern series, such as buying a phone followed by a cover.

3. Association Analysis

It analyses the set of items that generally occur together in a transactional dataset. It is also known as Market Basket Analysis for its wide use in retail sales. Two parameters are used for determining the association rules:

- It provides which identifies the common item set in the database.
- Confidence is the conditional probability that an item occurs when another item occurs in a transaction.

4. Classification

Classification is a data mining technique that categorizes items in a collection based on some predefined properties. It uses methods like if-then, decision trees or neural networks to predict a class or essentially classify a collection of items. A training set containing items whose properties are known is used to train the system to predict the category of items from an unknown collection of items.

5. Prediction

It defines predict some unavailable data values or spending trends. An object can be anticipated based on the attribute values of the object and attribute values of the classes. It can be a prediction of missing numerical values or increase or decrease trends in time-related information. There are primarily two types of predictions in data mining: numeric and class predictions.

- ***Numeric predictions*** are made by creating a linear regression model that is based on historical data. Prediction of numeric values helps businesses ramp up for a future event that might impact the business positively or negatively.
- ***Class predictions*** are used to fill in missing class information for products using a training data set where the class for products is known.

6. Cluster Analysis

In image processing, pattern recognition and bioinformatics, clustering is a popular data mining functionality. It is similar to classification, but the classes are not predefined. Data attributes represent the classes. Similar data are grouped together, with the difference being that a class label is not known. Clustering algorithms group data based on similar features and dissimilarities.

7. Outlier Analysis

Outlier analysis is important to understand the quality of data. If there are too many outliers, you cannot trust the data or draw patterns. An outlier analysis determines if there is something out of turn in the data and whether it indicates a situation that a business needs to consider and take measures to mitigate. An outlier analysis of the data that cannot be grouped into any classes by the algorithms is pulled up.

8. Evolution and Deviation Analysis

Evolution Analysis pertains to the study of data sets that change over time. Evolution analysis models are designed to capture evolutionary trends in data helping to characterize, classify, cluster or discriminate time-related data.

9. Correlation Analysis

Correlation is a mathematical technique for determining whether and how strongly two attributes is related to one another. It refers to the various types of data structures, such as trees and graphs, that can be combined with an item set or subsequence. It determines how well two numerically measured continuous variables are linked. Researchers can use this type of analysis to see if there are any possible correlations between variables in their study.

Types of Data Mining

If you haven't heard the term data mining yet, it would be good to have a little discussion about "data mining" before learning the types of data mining. In this article, we will learn the different types of data mining (or data mining methods). However, if you already know what data mining is, you can directly move on to data mining methods (or types).

What is Data Mining?

In General, data mining

is nothing but a process of finding or extracting useful information from huge volumes of data. You may get familiar if we use the term big data. Although using a big range of techniques can help us to use this information to increase revenues, cost-cutting and improve customer relationships, etc. It may be quite possible that you may be thinking that is why data mining is so important. The answer to this question is quite complex. However, it is not the answer that is actually big. You may have seen staggering numbers; the volumes of produced data are getting doubled every two years. However, this growth rate of the data is also increasing, or it will be correct to say that data is getting doubled even in less than two years.

Features of Data mining

These are the following key features that data mining usually allows us:

- Sift through all the chaotic and repetitive noise in your data.
- Allows understanding what is relevant and then making good use of that information to assess likely outcomes.

- Accelerate the pace of making informed decisions.

Why do we need Data Mining?

In today's modern world, we are all surrounded by big data, which is predicted to be grown by 40% by the next decade. You may wonder that the real fact is that we are drowning in the data, but at the same time, we are starving for knowledge (or useful Data). The main reason behind this, all this data creates noise which makes it difficult to mine. In short, we have generated tons of amorphous data but experiencing failing big data initiatives as the useful data is deeply buried inside. Therefore without powerful tools such as Data Mining, we cannot mine such data, and as a result, we will not get any benefits from that data.

Types of Data Mining

Each of the following data mining techniques serves several different business problems and provides a different insight into each of them. However, understanding the type of business problem you need to solve will also help in knowing which technique will be best to use, which will yield the best results. The Data Mining types can be divided into two basic parts that are as follows:

1. Predictive Data Mining Analysis
2. Descriptive Data Mining Analysis

1. Predictive Data Mining

As the name signifies, Predictive Data-Mining analysis works on the data that may help to know what may happen later (or in the future) in business. Predictive Data-Mining can also be further divided into four types that are listed below:

- Classification Analysis
- Regression Analysis
- Time Serious Analysis
- Prediction Analysis

2. Descriptive Data Mining

The main goal of the Descriptive Data Mining tasks is to summarize or turn given data into relevant information. The Descriptive Data-Mining Tasks can also be further divided into four types that are as follows:

- Clustering Analysis
- Summarization Analysis
- Association Rules Analysis
- Sequence Discovery Analysis

Here, we will discuss each of the data mining's types in detail. Below are several different data mining techniques that can help you find optimal outcomes as the results.

1. CLASSIFICATION ANALYSIS

This type of data mining technique is generally used in fetching or retrieving important and relevant information about the data & metadata. It is also even used to categories the different types of data format into different classes. If you focus on this article until it ends, you may definitely find out that Classification and clustering are similar data mining types. As clustering also categorizes or classify the data segments into the different data records known as the classes. However, unlike clustering, the data analyst would have the knowledge of different classes or clusters. Therefore in the classification analysis, you have to apply or implement the algorithms to decide in which way the new data should be categorized or classified. A classic example of classification analysis would be Outlook email. In Outlook, they use certain algorithms to characterize an email is legitimate or spam.

This technique is usually very helpful for retailers who can use it to study the buying habits of their different customers. Retailers can also study the past sales data and then lookout (or search) for products that customers usually buy together. After which, they can put those products nearby of each other in their retail stores to help customers save their time and as well as to increase their sales.

2. REGRESSION ANALYSIS

In statistical terms, regression analysis is a process usually used to identify and analyze the relationship among variables. It means one variable is dependent on another, but it is not vice versa. It is generally used for prediction and forecasting purposes. It can also help you understand the characteristic value of the dependent variable changes if any of the independent variables is varied.

3. Time Serious Analysis

A time series is a sequence of data points that are usually recorded at specific time intervals of points. Usually, they are - most often in regular time intervals (seconds, hours, days, months etc.). Almost every organization generates a high volume of data every day, such as sales figures, revenue, traffic, or operating cost. Time series data mining can help in generating valuable information for long-term business decisions, yet they are underutilized in most organizations.

4. Prediction Analysis

This technique is generally used to predict the relationship that exists between both the independent and dependent variables as well as the independent variables alone. It can also use to predict profit that can be achieved in future depending on the sale. Let us imagine that profit and sale are dependent and independent variables, respectively. Now, on the basis of what the past sales data says, we can make a profit prediction of the future using a regression curve.

5. Clustering Analysis

In Data Mining, this technique is used to create meaningful object clusters that contain the same characteristics. Usually, most people get confused with Classification, but they won't have any issues if they properly understand how both these techniques actually work. Unlike Classification that collects the objects into predefined classes, clustering stores objects in classes that are defined by it. To understand it in more detail, you can consider the following given example:

Example

Suppose you are in a library that is full of books on different topics. Now the real challenge for you is to organize those books so that readers don't face any problem finding out books on any particular topic. So here, we can use clustering to keep books with similarities in one particular shelf and then give those shelves a meaningful name or class. Therefore, whenever a reader looking for books on a particular topic can go straight to that shelf. Hence he won't be required to roam the entire library to find the book he wants to read.

6. SUMMARIZATION ANALYSIS

The Summarization analysis is used to store a group (or a set) of data in a more compact way and an easier-to-understand form. We can easily understand it with the help of an example:

Example

You might have used Summarization to create graphs or calculate averages from a given set (or group) of data. This is one of the most familiar and accessible forms of data mining.

7. ASSOCIATION RULE LEARNING

In general, it can be considered a method that can help us identify some interesting relations (dependency modeling) between different variables in large databases. This technique can also help us to unpack some hidden patterns in the data, which can be used to identify the variables within the data. It also helps in detecting the concurrence of different variables that appear very frequently in the dataset. Association rules are generally used for examining and forecasting the behavior of the customer. It is also highly recommended in the retail industry analysis. This technique is also used to determine shopping basket data analysis, catalogue design, product clustering, and store layout. In IT, programmers also uses the association rules to create programs capable of machine learning. Or in short, we can say that this data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.

8. Sequence Discovery Analysis

The primary goal of sequence discovery analysis is to discover interesting patterns in data on the basis of some subjective or objective measure of how interesting it is. Usually, this task involves discovering frequent sequential patterns with respect to a frequency support measure. Some people may often confuse it with time series as both the Sequence discovery analysis and Time series analysis contains the adjacent observation that are order dependent. However, if the people see both of them in a little more depth, their confusion can be easily avoided as the Time series analysis technique contains numerical data, whereas the Sequence discovery analysis contains discrete values or data.

Types of Sources of Data in Data Mining

In this post, we will discuss what are different sources of data that are used in data mining process. The data from multiple sources are integrated into a common source known as **Data Warehouse**.

Let's discuss what type of data can be mined:

1. ***Flat Files***
2. ***Relational Databases***
3. ***DataWarehouse***
4. ***Transactional Databases***
5. ***Multimedia Databases***
6. ***Spatial Databases***
7. ***Time Series Databases***
8. ***World Wide Web(WWW)***

1. ***Flat Files***
    - Flat files is defined as data files in text form or binary form with a structure that can be easily extracted by data mining algorithms.
    - Data stored in flat files have no relationship or path among themselves, like if a relational database is stored on flat file, then there will be no relations between the tables.
    - Flat files are represented by data dictionary. Eg: CSV file.
    - **Application**: Used in DataWarehousing to store data, Used in carrying data to and from server, etc.
2. ***Relational Databases***
    - A Relational database is defined as the collection of data organized in tables with rows and columns.
    - Physical schema in Relational databases is a schema which defines the structure of tables.
    - Logical schema in Relational databases is a schema which defines the relationship among tables.
    - Standard API of relational database is SQL.
    - **Application**: Data Mining, ROLAP model, etc.
3. ***DataWarehouse***
    - A datawarehouse is defined as the collection of data integrated from multiple sources that will queries and decision making.

- There are three types of datawarehouse: **Enterprise** datawarehouse, **Data Mart** and **Virtual** Warehouse.
- Two approaches can be used to update data in DataWarehouse: **Query-driven** Approach and **Update-driven** Approach.
- **Application**: Business decision making, Data mining, etc.

4. ***Transactional Databases***
   - Transactional databases is a collection of data organized by time stamps, date, etc to represent transaction in databases.
   - This type of database has the capability to roll back or undo its operation when a transaction is not completed or committed.
   - Highly flexible system where users can modify information without changing any sensitive information.
   - Follows ACID property of DBMS.
   - **Application**: Banking, Distributed systems, Object databases, etc.
5. ***Multimedia Databases***
   - Multimedia databases consists audio, video, images and text media.
   - They can be stored on Object-Oriented Databases.
   - They are used to store complex information in a pre-specified formats.
   - **Application**: Digital libraries, video-on demand, news-on demand, musical database, etc.
6. ***Spatial Database***
   - Store geographical information.
   - Stores data in the form of coordinates, topology, lines, polygons, etc.
   - **Application**: Maps, Global positioning, etc.
7. ***Time-series Databases***
   - Time series databases contains stock exchange data and user logged activities.
   - Handles array of numbers indexed by time, date, etc.
   - It requires real-time analysis.
   - **Application**: eXtremeDB, Graphite, InfluxDB, etc.
8. ***WWW***
   - WWW refers to World wide web is a collection of documents and resources like audio, video, text, etc which are identified by Uniform Resource Locators (URLs) through web browsers, linked by HTML pages, and accessible via the Internet network.
   - It is the most heterogeneous repository as it collects data from multiple resources.
   - It is dynamic in nature as Volume of data is continuously increasing and changing.
   - **Application**: Online shopping, Job search, Research, studying, etc.

Data Mining: Data Attributes and Quality

Data: It is how the data objects and their attributes are stored.

- An **attribute** is an object's property or characteristics. For example. A person's hair colour, air humidity etc.
- An attribute set defines an **object**. The **object** is also referred to as a record of the instances or entity.

Different **types of attributes or data types:**

1. **Nominal Attribute:**
   Nominal Attributes only provide enough attributes to differentiate between one object and another. Such as Student Roll No., Sex of the Person.
2. **Ordinal Attribute:**
   The ordinal attribute value provides sufficient information to order the objects. Such as Rankings, Grades, Height
3. **Binary Attribute:**
   These are 0 and 1. Where 0 is the absence of any features and 1 is the inclusion of any characteristics.
4. **Numeric attribute:**It is quantitative, such that quantity can be measured and represented in integer or real values ,are of two types
5. **Interval Scaled attribute:**
   It is measured on a scale of equal size units,these attributes allow us to compare such as temperature in C or F and thus values of attributes have ordered.
6. **Ratio Scaled attribute:** Both differences and ratios are significant for Ratio. For eg. age, length, and Weight.

Data Quality: Why do we pre-process the data?

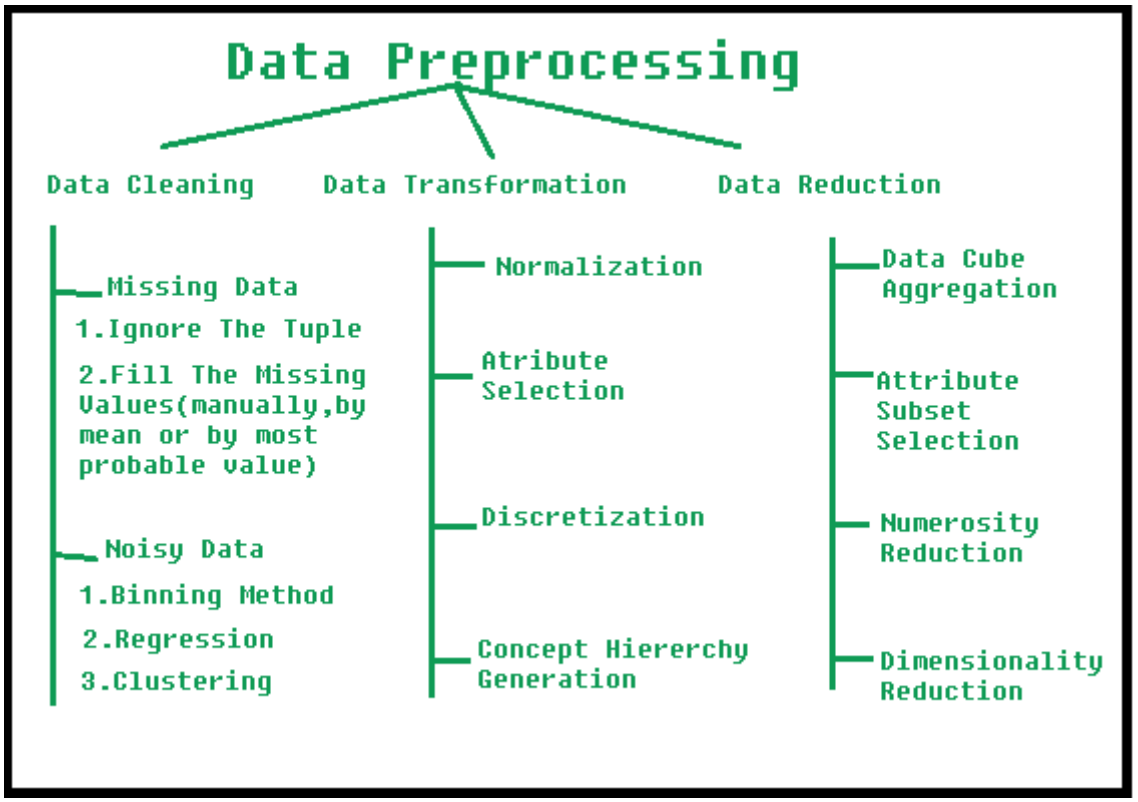
Many characteristics act as a deciding factor for data quality, such as incompleteness and incoherent information, which are common properties of the big database in the real world. Factors used for data quality assessment are:

- **Accuracy:**
  There are many possible reasons for flawed or inaccurate data here. i.e. Having incorrect values of properties that could be human or computer errors.
- **Completeness:**
  For some reasons, incomplete data can occur, attributes of interest such as customer information for sales & transaction data may not always be available.
- **Consistency:**
  Incorrect data can also result from inconsistencies in naming convention or data codes, or from input field incoherent format. Duplicate tuples need cleaning of details, too.
- **Timeliness:**
  It also affects the quality of the data. At the end of the month, several sales representatives fail to file their sales records on time. There are also several corrections & adjustments which flow into after the end of the month. Data stored in the database are incomplete for a time after each month.
- **Believability:**
  It is reflective of how much users trust the data.
- **Interpretability:**
  It is a reflection of how easy the users can understand the data.

Data Preprocessing in Data Mining

Preprocessing in Data Mining:

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



Steps Involved in Data Preprocessing:

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

(a). Missing Data: This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

1. **Ignore the tuples:** This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

2. **Fill the Missing values:** There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

- **(b). Noisy Data:** Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

1. **Binning Method:**
   This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

2. **Regression:**
   Here data can be made smooth by fitting it to a regression function.The regression

used may be linear (having one independent variable) or multiple (having multiple independent variables).

3. **Clustering:**
   This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation:

This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

2. **Attribute Selection:**
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

3. **Discretization:**
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

4. **Concept Hierarchy Generation:** Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute "city" can be converted to "country".

3. Data Reduction:

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.
The various steps to data reduction are:

1. **Data Cube Aggregation:** Aggregation operation is applied to data for the construction of the data cube.

2. **Attribute Subset Selection:** The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute. the attribute having p-value greater than significance level can be discarded.

3. **Numerosity Reduction:** This enable to store the model of data instead of whole data, for example: Regression Models.

4. **Dimensionality Reduction:** This reduce the size of data by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are: Wavelet transforms and PCA (Principal Component Analysis).

What are the measures of similarity and dissimilarity in data mining?

In data science, the similarity measure is a way of measuring how data samples are related or closed to each other. On the other hand, the dissimilarity measure is to tell how much the data objects are distinct.