DATA MINING & WAREHOUSING

ARCHANA G , Asst. Professor
CAROLINE ELIZHABETH.E, Asst. Professor,
SRINIVASAN COLLEGE OF ARTS AND SCIENCE

DATA MINING AND DATA WAREHOUSING

UNIT- I

Introduction to Data Mining:

Data mining is a process that is used by an organization to turn the raw data into useful data. Utilizing software to find patterns in large data sets, organizations can learn more about their customers to develop more efficient business strategies, boost sales, and reduce costs. Effective data collection, storage, and processing of the data are important advantages of data mining. Data mining method is been used to develop machine learning models.

What is Data Mining?

- It is basically the extraction of vital information/knowledge from a large set of data.
- Think of data as a large ground/rocky surface. We don't know what is inside it, we don't know if something useful is beneath the rocks.

Steps involved in Data Mining:

- Business Understanding ,
- Data Understanding,
- Data Preparation,
- Data Modeling,
- Evaluation and
- Deployment

Techniques used in Data Mining:

The techniques used in data mining are as listed below:

(i) Cluster Analysis,

It enables to identify a given user group according to common features in a database. These features could include age, geographic location, education level and so on.

(ii)Anomaly Detection,

It is used to determine when something is noticeably different from the regular pattern. It is used to eliminate any database inconsistencies or anomalies at the source.

(iii)Regression Analysis

This technique is used to make predictions based on relationships within the data set. For example, one can predict the stock rate of a particular product by analyzing the past rate and also by taking into account the different factors that determine the stock rate.

(iv)Classification

This deals with the things which have labels on it. Note in cluster detection, the things did not have a label in it and by using data mining we had to label and form into clusters, but in classification, there is information existing that can be easily classified using an algorithm.

Advantages of Data Mining:

Marketing/Retails

In order to create models, marketing companies use data mining. This was based on history to forecast who's going to respond to new marketing campaigns such as direct mail, online marketing, etc. This means that marketers can sell profitable products to targeted customers.

Finance/Banking

Data extraction provides information to financial institutions on loans and credit reports, data can determine good or bad credits by creating a model for historic customers. It also helps banks to detect fraudulent transactions by credit cards that protect the owner of a credit card.

Researchers

Data mining can motivate researchers to accelerate when the method analysis the data. Therefore they can work more time on other projects. Shopping behaviors can be detected. Most of the time, you may experience new problems while designing certain shopping patterns. Therefore data mining is used to solve these problems.

Determining Customer Groups

We are using data mining to respond from marketing campaigns to customers. It also provides information during the identification of customer groups.

Increases Brand Loyalty

In marketing campaigns, mining techniques are used. This is to understand their own customers ' needs and habits. And from that customers can also choose their brand's clothes. Thus, you can definitely be self-reliant with the help of this technique.

Helps in Decision Making

These data mining techniques are used by people to help them in making some sort of decisions in marketing or in business.

Increase Company Revenue

Data mining is a process in which some kind of technology is involved. One must collect information on goods sold online, this eventually reduces product costs and services, which is one of the benefits of data mining.

To Predict Future Trends

All information factors are part of the working nature of the system. The data mining systems can also be obtained from these. They can help you predict future trends and with the help of this technology, this is quite possible. And people also adopt behavioral changes.

Data Mining Algorithms:

- ### The k-means Algorithm

  This algorithm is a simple method of partitioning a given data set into the user-specified number of clusters.

  This algorithm works on d-dimensional vectors, D={xi | i= 1, … N} where i is the data point. To get these initial data seeds, the data has to be sampled at random. This sets the solution of clustering a small subset of data, the global mean of data k times.

  This algorithm can be paired with another algorithm to describe non-convex clusters. It creates k groups from the given set of objects.

  It explores the entire data set with its cluster analysis. It is simple and faster than other algorithms when it is used with other algorithms. This algorithm is mostly classified as semi-supervised.

- ### Naive Bayes Algorithm

  This algorithm is based on Bayes theorem. This algorithm is mainly used when the dimensionality of inputs is high.

This classifier can easily calculate the next possible output. New raw data can be added during the runtime and it provides a better probabilistic classifier.

Each class has a known set of vectors which aim at creating a rule which allows the objects to be assigned to classes in the future.

The vectors of variables describe the future objects. This is one of the easiest algorithms as it is easy to construct and does not have any complicated parameter estimation schemas.

- Support Vector Machines Algorithm

It is formed on the basis of structural risk minimization and statistical learning theory. The decision boundaries must be identified which is known as a hyperplane.

It helps in the optimal separation of classes. The main job of SVM is to identify the maximizing the margin between two classes.

The margin is defined as the amount of space between two classes. A hyperplane function is like an equation for the line, y= MX + b. SVM can be extended to perform numerical calculations as well.

- The Apriori Algorithm

Join: The whole database is used for the hoe frequent 1 item sets.

Prune: This item set must satisfy the support and confidence to move to the next round for the 2 item sets.

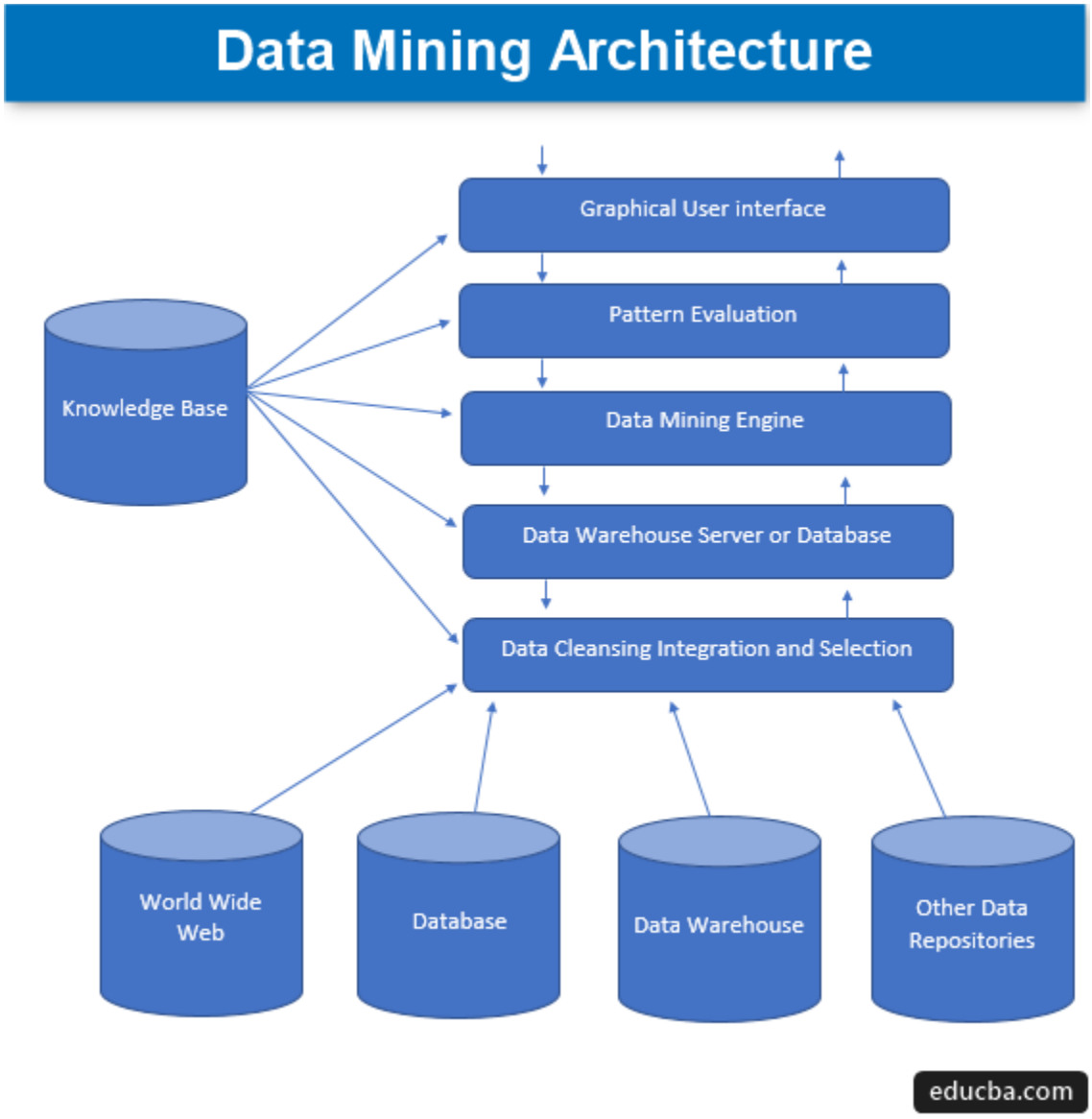
Repeat: Until the pre-defined size is not reached till then this is repeated for each itemset level.

Data Mining Applications

- Financial firms, banks, and their analysis
- Health care domain and insurance domain
- Application in the domain of transportation
- Applications of data mining in the field of medicine
- Education
- Manufacturing Engineering

Data Mining Architecture:

The data mining is the technique of extracting interesting knowledge from a set of huge amounts of data which then is stored in many data sources such as file systems, data warehouses, databases.



Data Sources

A huge variety of present documents such as data warehouse, database, www or popularly called a World wide web which becomes the actual data sources.

Most of the times, it can also be the case that the data is not present in any of these golden sources but only in the form of text files, plain files or sequence files or spreadsheets and then the data needs to be processed in a very similar way as the processing would be done upon the data received from golden sources.

Data Warehouse Server or Database

The database server is the actual space where the data is contained once it is received from the various number of data sources.

The server contains the actual set of data which becomes ready to be processed and therefore the server manages the data retrieval. All this activity is based on the request for data mining of the person.

Data Mining Engine

In the case of data mining, the engine forms the core component and is the most vital part, or to say the driving force which handles all the requests and manages them and is used to contain a number of modules.

The number of modules present includes mining tasks such as classification technique, association technique, regression technique, characterization, prediction and clustering, time series analysis, naive Bayes, support vector machines, ensemble methods, boosting and bagging techniques, random forests, decision trees, etc.

Graphical User Interface

When the data is communicated with the engines and among various pattern evaluation of modules, it becomes a necessity to interact with the various components present and make it more user friendly,

so that the efficient and effective use of all the present components could be made and therefore arises the need of a graphical user interface popularly known as GUI.

Knowledge Base

This is the component that forms the base of the overall data mining process as it helps in guiding the search or in the evaluation of interestingness of the patterns formed. This knowledgebase consists of user beliefs and also the data obtained from user

experiences which are in turn helpful in the data mining process. The engine might get its set of inputs from the created knowledge base and thereby provides more efficient, accurate and reliable results.

Advantages of Data Mining Process

The advantage of data mining includes not only the ones related to business but also ones like medicine, weather forecast, healthcare, transportation, insurance, government, etc. Some of the advantages include:

Marketing/Retail: It helps all the marketing companies and firms to build models which are based on a historic set of data and information in order to predict the responsiveness to the marketing campaigns prevailing today such as online marketing campaign, direct mail, etc.

Finance/Banking: The data mining involves financial institutions provide information about loans and also credit reporting. When the model is built on historical information, good or bad loans can then be determined by the financial institutions. Also, fraudulent and suspicious transactions are also monitored by the banks.

Manufacturing: The faulty equipment and the quality of the manufactured products can be determined by making use of the optimal parameters for controlling. For example, for some of the semi-conductor development industries, water hardness and quality become a major challenge as it tends to affect the quality of produce of their product.

Government: The governments can be benefitted with the monitoring and gauging the suspicious activities to avoid anti-money laundering activities.

Different Stages of Data Mining Process

Data cleansing: This is a very initial stage in the case of data mining where the classification of the data becomes an essential component to obtain final data analysis. It involves identifying and removal of inaccurate and tricky data from a set of tables, database, and recordset. Some techniques include the ignorance of tuple which is mainly found when the class label is not in place, the next technique requires filling of the missing values on its own, replacement of missing values and incorrect values with global constants or predictable or mean values.

Data integration: It is a technique which involves the merging of the new set of information with the existing set. The source may, however, involve many data sets, databases or flat files. The customary implementation for data integration is the creation of an EDW (enterprise data warehouse) which then talks about two concepts- tight as well as loose coupling, but let's not dig into the detail.

Data transformation: This requires the transformation of data within formats generally from the source system to the required destination system. Some strategies include Smoothing, Aggregation, Normalization, Generalization and attribute construction.

Data discretization: The techniques which can split the domain of continuous attribute along intervals is called data discretization wherein the datasets are stored in small chunks and thereby making our study much more efficient. Two strategies involve Top-down discretization and bottom-up discretization.

Concept hierarchies: They minimize the data by replacing and collecting low-level concepts from high-level concepts. The multi-dimensional data with multiple levels of abstraction are defined by concept hierarchies. The methods are Binning, histogram analysis, cluster analysis, etc.

Pattern evaluation and data presentation: If the data is presented in an efficient manner, the client, as well as the customers, can make use of it in the best possible way. After going through the above set of stages the data then is presented in forms of graphs and diagrams and thereby understanding it with minimum statistical knowledge.

Association Rules Mining:

Association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Piatetsky-Shapiro describes analyzing and presenting strong rules discovered in databases using different measures of interestingness.

Based on the concept of strong rules, Rakesh Agrawal et al. introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets

For example, the rule found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements.

Definition

Example database with 4 items and 5 transactions

transaction ID	milk	bread	butter	beer
1	1	1	0	0
2	0	0	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Uses of association rules in data mining

In data mining, association rules are useful for analyzing and predicting customer behavior. They play an important part in customer analytics, market basket analysis, product clustering, catalog design and store layout.

Programmers use association rules to build programs capable of machine learning. Machine learning is a type of artificial intelligence (AI) that seeks to build programs with the ability to become more efficient without being explicitly programmed.

UNIT- II

Classifications:

It is a Data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts.

Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known.
Example: Before starting any Project, we need to check it's feasibility. In this case, a classifier is required to predict class labels such as 'Safe' and 'Risky' for adopting the Project and to further approve it. It is a two-step process such as :

Learning Step (Training Phase): Construction of Classification Model
Different Algorithms are used to build a classifier by making the model learn using the training set available. The model has to be trained for the prediction of accurate results.

Classification Step: Model used to predict class labels and testing the constructed model on test data and hence estimate the accuracy of the classification rules.

Training and Testing: Suppose there is a person who is sitting under a fan and the fan starts falling on him, he should get aside in order not to get hurt. So, this is his training part to move away. While Testing if the person sees any heavy object coming towards him or falling on him and moves aside then

the system is tested positively and if the person does not move aside then the system is negatively tested.
Same is the case with the data, it should be trained in order to get the accurate and best results.

There are certain data types associated with data mining that actually tells us the format of the file (whether it is in text format or in numerical format).

Attributes – Represents different features of an object. Different types of attributes are:

1. Binary: Possesses only two values i.e. True or False
   Example: Suppose there is a survey of evaluating some product. We need to check whether it's useful or not. So, the Customer has to answer it in Yes or No.
   Product usefulness: Yes / No
   - Symmetric: Both values are equally important in all aspects
   - Asymmetric: When both the values may not be important.

2. Nominal: When more than two outcomes are possible. It is in Alphabet form rather than being in Integer form.
   Example: One needs to choose some material but of different colors. So, the color might be Yellow, Green, Black, Red.
   Different Colors: Red, Green, Black, Yellow

- **Ordinal**: Values that must have some meaningful order.
  Example: Suppose there are grade sheets of few students which might contain different grades as per their performance such as A, B, C, D
  Grades: A, B, C, D

  - **Continuous**: May have infinite number of values, it is in float type
    Example: Measuring weight of few Students in a sequence or orderly manner i.e. 50, 51, 52, 53
    Weight: 50, 51, 52, 53
  - Discrete: Finite number of values.
    Example: Marks of a Student in a few subjects: 65, 70, 75, 80, 90
    Marks: 65, 70, 75, 80, 90

Syntax:

Mathematical Notation: Classification is based on building a function taking input feature vector "X" and predicting its outcome "Y" (Qualitative response taking values in set C)

Here Classifier (or model) is used which is a Supervised function, can be designed manually based on expert's knowledge. It has been constructed to predict class labels (Example: Label – "Yes" or "No" for the approval of some event).

Classifiers can be categorized into two major types:

1. **Discriminative**: It is a very basic classifier and determines just one class for each row of data. It tries to model just by depending on the observed data, depends heavily on the quality of data rather than on distributions.
   Example: Logistic Regression

2. **Generative**: It models the distribution of individual classes and tries to learn the model that generates the data behind the scenes by estimating assumptions and distributions of the model. Used to predict the unseen data.
   Example: Naive Bayes Classifier

Classifiers Of Machine Learning:

1. Decision Trees
2. Bayesian Classifiers
3. Neural Networks
4. K-Nearest Neighbour
5. Support Vector Machines
6. Linear Regression
7. Logistic Regression

Advantages:

- Mining Based Methods are cost effective and efficient
- Helps in identifying criminal suspects
- Helps in predicting risk of diseases
- Helps Banks and Financial Institutions to identify defaulters so that they may approve Cards, Loan, etc.

Disadvantages:

Privacy: When the data is either are chances that a company may give some information about their customers to other vendors or use this information for their profit.

Accuracy Problem: Selection of Accurate model must be there in order to get the best accuracy and result.

APPLICATIONS:

- Marketing and Retailing
- Manufacturing
- Telecommunication Industry
- Intrusion Detection
- Education System
- Fraud Detection

What is Clustering?

- Clustering is the process of making a group of abstract objects into classes of similar objects.

Points to Remember

- A cluster of data objects can be treated as one group.
- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
- The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

Applications of Cluster Analysis

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.

- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

Requirements of Clustering in Data Mining

The following points throw light on why clustering is required in data mining −

- **Scalability** − We need highly scalable clustering algorithms to deal with large databases.
- **Ability to deal with different kinds of attributes** − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- **Discovery of clusters with attribute shape** − The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- **High dimensionality** − The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
- **Ability to deal with noisy data** − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
- **Interpretability** − The clustering results should be interpretable, comprehensible, and usable.

Clustering Methods

Clustering methods can be classified into the following categories −

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

Partitioning Method

Suppose we are given a database of ‘n’ objects and the partitioning method constructs ‘k’ partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements −

- Each group contains at least one object.
- Each object must belong to exactly one group.

Points to remember −

- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

Hierarchical Methods

This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here −

- Agglomerative Approach
- Divisive Approach

Agglomerative Approach

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

Divisive Approach

This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

Approaches to Improve Quality of Hierarchical Clustering

Here are the two approaches that are used to improve the quality of hierarchical clustering −

- Perform careful analysis of object linkages at each hierarchical partitioning.
- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

Density-based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantages

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

Model-based methods

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

Constraint-based Method

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

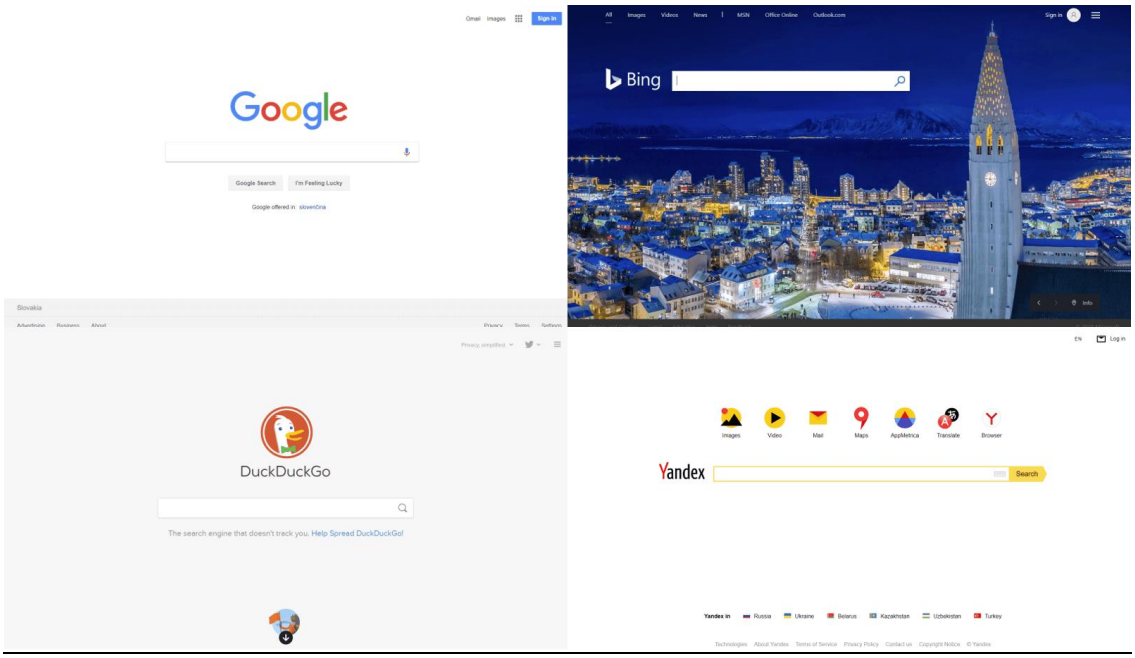
UNIT- III

Search Engine

A search engine is an online tool that searches for results in its database based on the search query (keyword) submitted by the internet user. The results are usually websites that semantically match with the search query.

Search engines find the results in their database, sort them and make an ordered list of these results based on the search algorithm. This list is generally called the search engine results page

There are many search engines on the market, while the most widely used is Google. Many website browsers such as Chrome, Firefox, Safari or Edge usually come with a default search engine set as a home page or starting page.



How search engines work

There may be some differences in how the search engines work but the fundamentals remain the same. Each of them has to do the following tasks:

1. Crawling
2. Indexing
3. Creating results

1. Crawling

Search engines have their own crawlers, small bots that scan websites on the world wide web. These little bots scan all sections, folders, subpages, content, everything they can find on the website.

Crawling is based on finding hypertext links that refer to other websites. By parsing these links, the bots are able to recursively find new sources to crawl.

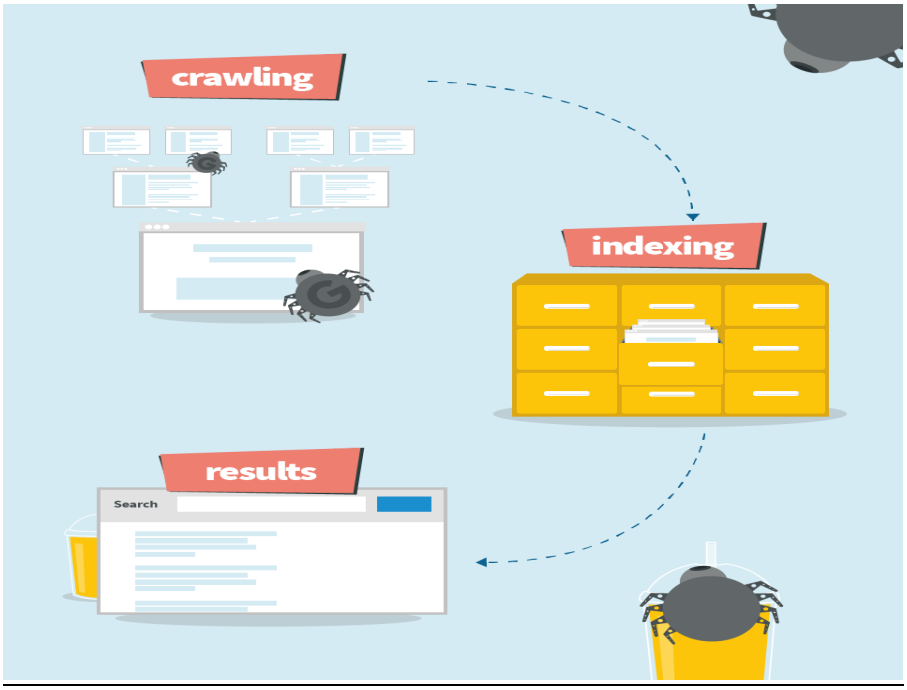
2. Indexing

Once the bots crawl the data, it’s time for indexing. The index is basically an online library of websites.

Your website has to be indexed in order to be displayed in the search engine results page. Keep in mind that indexing is a constant process. Crawlers come back to each website to detect new data.

3. Creating results

Search engines create the results once the user submits a search query. It's a process of checking the query against all website records in the index. Based on the algorithm, the search engine picks the best results and creates an ordered list.



What is a search engine algorithm?

A search engine algorithm is a unique formula that determines how the websites are sorted in the search engine results page. It's a trademark of each search engine, therefore, it's kept secret.

The algorithm is a mixture of various ranking factors. You will find many articles dealing with the "real" Google ranking factors. The truth is that even when you know what the factors are, you don't know their exact weights.

The most important ranking factor of all search engines should be the *relevancy*. The main goal of search engines is to find what the internet user is looking for.

When it comes to Google, the major ranking factors are:

- Website/content relevancy
- Strength and relevancy of backlinks

Some of the other very important ranking factors are:

- Mobile optimization
- Content structure and optimization
- Usability
- Accessibility
- Page speed
- Social signals
- Overall domain authority

The most popular search engines

In terms of worldwide popularity, Google has been #1 for years. This is the list of top 10 most popular search engines:

1. Google

Google is the giant in the industry and has probably the most sophisticated algorithm. It includes machine learning, AI and RankBrain, another algorithm that can tweak the weights of ranking factors according to user's behavior and quality of previous results. More than 70% internet users worldwide do their searches on Google since 1998.

2. Baidu

Baidu is the first search engine launched in China in 2000. It's like Chinese Google. Baidu cooperates with companies such as Microsoft, Qualcomm, Intel or Daimler on various AI projects. Similarly to Google, they offer a lot of other solutions such as cloud services, maps, social networking, image and video search and many others.

3. Bing

Microsoft launched their search engine in 2009 as a new project after earlier search engines MSN Search and Windows Live Search. The main goal was to develop a competitor for Google. From the global point of view, it's not really there but in the US, Bing is the 2nd most popular tool for the internet searches.

4. Yahoo!

Originally, it was one of the most widely used email providers and search engines. The company grew significantly in the 1990s but after 2000, they somehow started lacking the innovation and lost their value. In 2017, Yahoo! was acquired by Verizon Communications.

5. Yandex

Yandex Search is the major Russian search engine. According to Wikipedia, Yandex generates more than 50% of all searches in Russia. Though the algorithm is not as sophisticated as Google, it constantly gets better by integrating AI and machine learning that analyze searches and learn from them.

6. Ask

Ask (formerly Ask Jeeves) was launched in 1996. It was designed to answer questions submitted to the search form. Thanks to the Ask toolbar, this search engine was able to compete with big players such as Bing, Yahoo! and Google. Unfortunately, the toolbar was many times installed as an unwanted browser feature.

7. DuckDuckGo

DuckDuckGo is a bit different search engine. They protect the users' privacy by not tracking any information. DuckDuckGo doesn't show personalized results based on your previous searches. Likewise, advertisers can't follow the behavior of the users. On the other hand, you can launch ads via Bing since DuckDuckGo is their search partner together with Yahoo.

8. Naver

Naver is the Google of South Korea. This search engine covers around 75% of searches in the country. It was launched in 1999 and in 2000 it was able to pull out various types of results that match the entered keywords. The results included

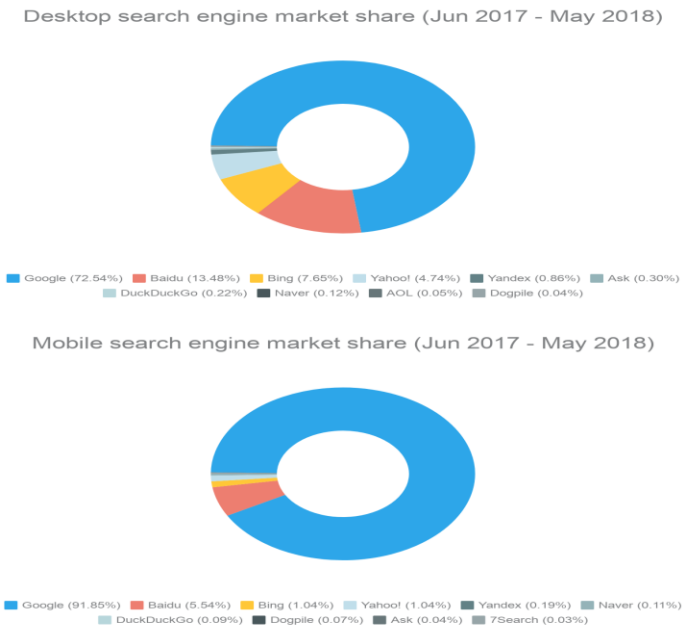
websites, images, blogs, restaurants, shops, etc. Google launched this feature 5 years later.

9. AOL

In the 1990's, AOL was one of the biggest crawler-based search engines. As a company, AOL offers a lot of other services: email service, instant messenger, video content, yellow pages, city guides. The AOL Search is nowadays used by not more than 0.5% of the internet users.

10. Dogpile

Dogpile is a metasearch engine, so it creates the search engine results page by doing simultaneous search requests for the same query in other search engines, namely: Google, Yahoo!, Yandex and others. Dogpile was launched in 1995.



Difference between Search Engine and Web Browser:

SEARCH ENGINE	WEB BROWSER
A search engine is used to find the information in the World Wide Web and displays the results at one place.	Web Browser uses the search engine to retrieve and view the information from web pages present on the web servers.

SEARCH ENGINE	WEB BROWSER
Search engine is intended to gather Information regarding several URL's and to maintain it.	Web Browsers are intended to Display the web page of the current URL available at the server.
A search engine contains its own database	No database is required in Web browser. It contains only cache memory to store cookies.
Example of famous search engines are: Google, Yahoo, Bing, DuckDuckgo, Baidu Internet Explorer.	Some of the widely used web browsers are: Mozilla Firefox, Netscape Navigator, and Google Chrome.

Unit--IV

Data Warehousing

A ***Data Warehousing*** (DW) is process for collecting and managing data from varied sources to provide meaningful business insights.

- A Data warehouse is typically used to connect and analyze business data from heterogeneous sources. The data warehouse is the core of the BI system which is built for data analysis and reporting.

- It is a blend of technologies and components which aids the strategic use of data. It is electronic storage of a large amount of information by a business which is designed for query and analysis instead of transaction processing.

- The decision support database (Data Warehouse) is maintained separately from the organization's operational database. However, the data warehouse is not a product but an environment. It is an architectural construct of an information system which provides users with current and historical decision support information which is difficult to access or present in the traditional operational data store.

- You many know that a 3NF-designed database for an inventory system many have tables related to each other.

- For example, a report on current inventory information can include more than 12 joined conditions. This can quickly slow down the response time of the query and report. A data warehouse provides a new design which can help to reduce the response time and helps to enhance the performance of queries for reports and analytics.

Data warehouse system is also known by the following name:

- Decision Support System (DSS)
- Executive Information System
- Management Information System
- Business Intelligence Solution
- Analytic Application
- Data Warehouse

History of Datawarehouse

The Datawarehouse benefits users to understand and enhance their organization's performance. The need to warehouse data evolved as computer systems became more complex and needed to handle increasing amounts of Information. However, Data Warehousing is a not a new thing.

Here are some key events in evolution of Data Warehouse-

- 1960- Dartmouth and General Mills in a joint research project, develop the terms dimensions and facts.
- 1970- A Nielsen and IRI introduces dimensional data marts for retail sales.
- 1983- Tera Data Corporation introduces a database management system which is specifically designed for decision support
- Data warehousing started in the late 1980s when IBM worker Paul Murphy and Barry Devlin developed the Business Data Warehouse.
- However, the real concept was given by Inmon Bill. He was considered as a father of data warehouse. He had written about a variety of topics for building, usage, and maintenance of the warehouse & the Corporate Information Factory.

How Datawarehouse works?

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

Data may be:

1. Structured
2. Semi-structured
3. Unstructured data

The data is processed, transformed, and ingested so that users can access the processed data in the Data Warehouse through Business Intelligence tools, SQL clients, and spreadsheets. A data warehouse merges information coming from different sources into one comprehensive database.

By merging all of this information in one place, an organization can analyze its customers more holistically. This helps to ensure that it has considered all the information available. Data warehousing makes data mining possible. Data mining is looking for patterns in the data that may lead to higher sales and profits.

Types of Data Warehouse

Three main types of Data Warehouses are:

1. Enterprise Data Warehouse:

Enterprise Data Warehouse is a centralized warehouse. It provides decision support service across the enterprise. It offers a unified approach for organizing and representing data. It also provide the ability to classify data according to the subject and give access according to those divisions.

2. Operational Data Store:

Operational Data Store, which is also called ODS, are nothing but data store required when neither Data warehouse nor OLTP systems support organizations reporting needs. In ODS, Data warehouse is refreshed in real time. Hence, it is widely preferred for routine activities like storing records of the Employees.

3. Data Mart:

A data mart is a subset of the data warehouse. It specially designed for a particular line of business, such as sales, finance, sales or finance. In an independent data mart, data can collect directly from sources.

General stages of Data Warehouse

Earlier, organizations started relatively simple use of data warehousing. However, over time, more sophisticated use of data warehousing begun.

The following are general stages of use of the data warehouse:

Offline Operational Database:

In this stage, data is just copied from an operational system to another server. In this way, loading, processing, and reporting of the copied data do not impact the operational system's performance.

Offline Data Warehouse:

Data in the Datawarehouse is regularly updated from the Operational Database. The data in Datawarehouse is mapped and transformed to meet the Datawarehouse objectives.

Real time Data Warehouse:

In this stage, Data warehouses are updated whenever any transaction takes place in operational database. For example, Airline or railway booking system.

Integrated Data Warehouse:

In this stage, Data Warehouses are updated continuously when the operational system performs a transaction. The Datawarehouse then generates transactions which are passed back to the operational system.

Components of Data warehouse

Four components of Data Warehouses are:

Load manager: Load manager is also called the front component. It performs with all the operations associated with the extraction and load of data into the warehouse. These operations include transformations to prepare the data for entering into the Data warehouse.

Warehouse Manager: Warehouse manager performs operations associated with the management of the data in the warehouse. It performs operations like analysis of data to ensure consistency, creation of indexes and views, generation of denormalization and aggregations, transformation and merging of source data and archiving and baking-up data.

Query Manager: Query manager is also known as backend component. It performs all the operation operations related to the management of user queries. The operations of this Data warehouse components are direct queries to the appropriate tables for scheduling the execution of queries.

End-user access tools:

This is categorized into five different groups like 1. Data Reporting 2. Query Tools 3. Application development tools 4. EIS tools, 5. OLAP tools and data mining tools.

Who needs Data warehouse?

Data warehouse is needed for all types of users like:

- Decision makers who rely on mass amount of data
- Users who use customized, complex processes to obtain information from multiple data sources.
- It is also used by the people who want simple technology to access the data
- It also essential for those people who want a systematic approach for making decisions.
- If the user wants fast performance on a huge amount of data which is a necessity for reports, grids or charts, then Data warehouse proves useful.
    - Data warehouse is a first step If you want to discover 'hidden patterns' of data-flows and groupings.
    -

What Is a Data Warehouse Used For?

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

Airline:

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

Banking:

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

Healthcare:

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

Public sector:

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

Investment and Insurance sector:

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

Retain chain:

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

Telecommunication:

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

Hospitality Industry:

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

Steps to Implement Data Warehouse

The best way to address the business risk associated with a Datawarehouse implementation is to employ a three-prong strategy as below

1. ***Enterprise strategy***: Here we identify technical including current architecture and tools. We also identify facts, dimensions, and attributes. Data mapping and transformation is also passed.
2. ***Phased delivery***: Datawarehouse implementation should be phased based on subject areas. Related business entities like booking and billing should be first implemented and then integrated with each other.
3. ***Iterative Prototyping***: Rather than a big bang approach to implementation, the Datawarehouse should be developed and tested iteratively.

Here, are key steps in Datawarehouse implementation along with its deliverables.

Step	*Tasks*	*Deliverables*
1	Need to define project scope	Scope Definition
2	Need to determine business needs	Logical Data Model
3	Define Operational Datastore requirements	Operational Data Store Model
4	Acquire or develop Extraction tools	Extract tools and Software
5	Define Data Warehouse Data requirements	Transition Data Model
6	Document missing data	To Do Project List
7	Maps Operational Data Store to Data Warehouse	D/W Data Integration Map
8	Develop Data Warehouse Database design	D/W Database Design
9	Extract Data from Operational Data Store	Integrated D/W Data Extracts
10	Load Data Warehouse	Initial Data Load
11	Maintain Data Warehouse	On-going Data Access and Subsequent Loads

Best practices to implement a Data Warehouse

- Decide a plan to test the consistency, accuracy, and integrity of the data.
- The data warehouse must be well integrated, well defined and time stamped.
- While designing Datawarehouse make sure you use right tool, stick to life cycle, take care about data conflicts and ready to learn you're your mistakes.
- Never replace operational systems and reports
- Don't spend too much time on extracting, cleaning and loading data.

- Ensure to involve all stakeholders including business personnel in Datawarehouse implementation process. Establish that Data warehousing is a joint/ team project. You don't want to create Data warehouse that is not useful to the end users.
- Prepare a training plan for the end users.

Why We Need Data Warehouse? Advantages & Disadvantages

Advantages of Data Warehouse:

- Data warehouse allows business users to quickly access critical data from some sources all in one place.
- Data warehouse provides consistent information on various cross-functional activities. It is also supporting ad-hoc reporting and query.
- Data Warehouse helps to integrate many sources of data to reduce stress on the production system.
- Data warehouse helps to reduce total turnaround time for analysis and reporting.
- Restructuring and Integration make it easier for the user to use for reporting and analysis.
- Data warehouse allows users to access critical data from the number of sources in a single place. Therefore, it saves user's time of retrieving data from multiple sources.
- Data warehouse stores a large amount of historical data. This helps users to analyze different time periods and trends to make future predictions.

Disadvantages of Data Warehouse:

- Not an ideal option for unstructured data.
- Creation and Implementation of Data Warehouse is surely time confusing affair.
- Data Warehouse can be outdated relatively quickly
- Difficult to make changes in data types and ranges, data source schema, indexes, and queries.
- The data warehouse may seem easy, but actually, it is too complex for the average users.

- Despite best efforts at project management, data warehousing project scope will always increase.
- Sometime warehouse users will develop different business rules.
- Organisations need to spend lots of their resources for training and Implementation purpose.

The Future of Data Warehousing

- Change in ***Regulatory constrains*** may limit the ability to combine source of disparate data. These disparate sources may include unstructured data which is difficult to store.
- As the ***size*** of the databases grows, the estimates of what constitutes a very large database continue to grow. It is complex to build and run data warehouse systems which are always increasing in size. The hardware and software resources are available today do not allow to keep a large amount of data online.
- ***Multimedia data*** cannot be easily manipulated as text data, whereas textual information can be retrieved by the relational software available today. This could be a research subject.

Data Warehouse Tools

There are many Data Warehousing tools are available in the market. Here, are some most prominent one:

1. MarkLogic:

MarkLogic is useful data warehousing solution that makes data integration easier and faster using an array of enterprise features. This tool helps to perform very complex search operations. It can query different types of data like documents, relationships, and metadata.

https://developer.marklogic.com/products/

2. Oracle:

Oracle is the industry-leading database. It offers a wide range of choice of data warehouse solutions for both on-premises and in the cloud. It helps to optimize customer experiences by increasing operational efficiency.

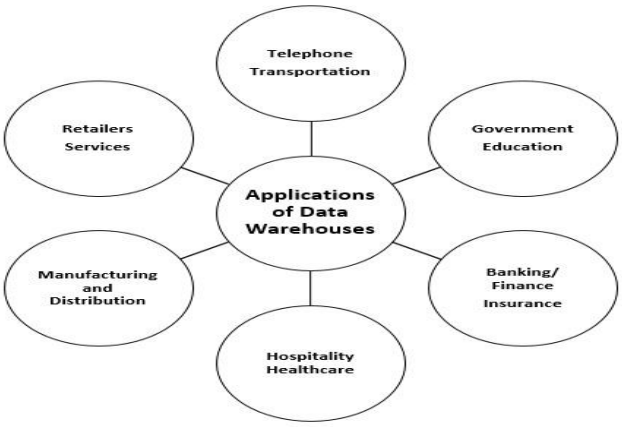
https://www.oracle.com/index.html

3. Amazon RedShift:

Amazon Redshift is Data warehouse tool. It is a simple and cost-effective tool to analyze all types of data using standard SQL and existing BI tools. It also allows running complex queries against petabytes of structured data, using the technique of query optimization.

https://aws.amazon.com/redshift/?nc2=h_m1

Applications of data warehouse



Banking Industry

In the banking industry, concentration is given to risk management and policy reversal as well analyzing consumer data, market trends, government regulations and reports, and more importantly financial decision making.

Most banks also use warehouses to manage the resources available on deck in an effective manner. Certain banking sectors utilize them for market research, performance analysis of each product, interchange and exchange rates, and to develop marketing programs.

Analysis of card holder's transactions, spending patterns and merchant classification, all of which provide the bank with an opportunity to introduce special offers and lucrative deals based on cardholder activity. Apart from all these, there is also scope for co-branding.

Finance Industry

Similar to the applications seen in banking, mainly revolve around evaluation and trends of customer expenses which aids in maximizing the profits earned by their clients.

Consumer Goods Industry

They are used for prediction of consumer trends, inventory management, market and advertising research. In-depth analysis of sales and production is also carried out. Apart from these, information is exchanged business partners and clientele.

Government and Education

The federal government utilizes the warehouses for research in compliance, whereas the state government uses it for services related to human resources like recruitment, and accounting like payroll management.

The government uses data warehouses to maintain and analyze tax records, health policy records and their respective providers, and also their entire criminal law database is connected to the state's data warehouse. Criminal activity is predicted from the patterns and trends, results of the analysis of historical data associated with past criminals.

Universities use warehouses for extracting of information used for the proposal of research grants, understanding their student demographics, and human resource management. The entire financial department of most universities depends on data warehouses, inclusive of the Financial Aid department.

Healthcare

One of the most important sector which utilizes data warehouses is the Healthcare sector. All of their financial, clinical, and employee records are fed to warehouses as it helps them to strategize and predict outcomes, track and analyze their service feedback, generate patient reports, share data with tie-in insurance companies, medical aid services, etc.

Hospitality Industry

A major proportion of this industry is dominated by hotel and restaurant services, car rental services, and holiday home services. They utilize warehouse services to design and evaluate their advertising and promotion campaigns where they target customers based on their feedback and travel patterns.

Insurance

As the saying goes in the insurance services sector, “Insurance can never be bought, it can be only be sold”, the warehouses are primarily used to analyze data patterns and customer trends, apart from maintaining records of already existing participants. The design of tailor-made customer offers and promotions is also possible through warehouses.

Manufacturing and Distribution Industry

This industry is one of the most important sources of income for any state. A manufacturing organization has to take several make-or-buy decisions which can influence the future of the sector, which is why they utilize high-end OLAP tools as a part of data warehouses to predict market changes, analyze current business trends, detect warning conditions, view marketing developments, and ultimately take better decisions.

They also use them for product shipment records, records of product portfolios, identify profitable product lines, analyze previous data and customer feedback to evaluate the weaker product lines and eliminate them.

For the distributions, the supply chain management of products operates through data warehouses.

The Retailers

Retailers serve as middlemen between producers and consumers. It is important for them to maintain records of both the parties to ensure their existence in the market.

They use warehouses to track items, their advertising promotions, and the consumers buying trends. They also analyze sales to determine fast selling and slow selling product lines and determine their shelf space through a process of elimination.

Services Sector

Data warehouses find themselves to be of use in the service sector for maintenance of financial records, revenue patterns, customer profiling, resource management, and human resources.

Telephone Industry

The telephone industry operates over both offline and online data burdening them with a lot of historical data which has to be consolidated and integrated.

Apart from those operations, analysis of fixed assets, analysis of customer's calling patterns for sales representatives to push advertising campaigns, and tracking of customer queries, all require the facilities of a data warehouse.

Transportation Industry

In the transportation industry, data warehouses record customer data enabling traders to experiment with target marketing where the marketing campaigns are designed by keeping customer requirements in mind.

The internal environment of the industry uses them to analyze customer feedback, performance, manage crews on board as well as analyze customer financial reports for pricing strategies.

Architecture of data warehouse

Three-Tier Data Warehouse Architecture

Generally a data warehouses adopts a three-tier architecture. Following are the three tiers of the data warehouse architecture.

- **Bottom Tier** − The bottom tier of the architecture is the data warehouse database server. It is the relational database system. We use the back end tools and utilities to feed data into the bottom tier. These back end tools and utilities perform the Extract, Clean, Load, and refresh functions.

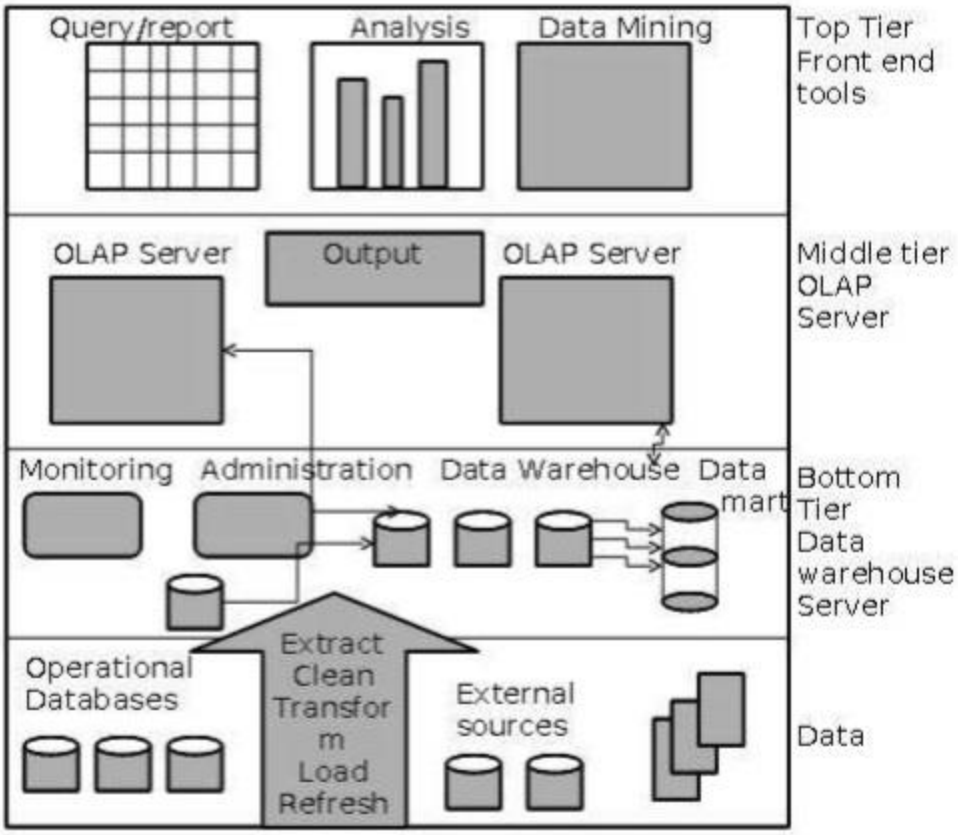
- **Middle Tier** − In the middle tier, we have the OLAP Server that can be implemented in either of the following ways.

  - By Relational OLAP (ROLAP), which is an extended relational database management system. The ROLAP maps the operations on multidimensional data to standard relational operations.

  - By Multidimensional OLAP (MOLAP) model, which directly implements the multidimensional data and operations.

- **Top-Tier** − This tier is the front-end client layer. This layer holds the query tools and reporting tools, analysis tools and data mining tools.

The following diagram depicts the three-tier architecture of data warehouse −



Data Warehouse Models

From the perspective of data warehouse architecture, we have the following data warehouse models −

- Virtual Warehouse
- Data mart
- Enterprise Warehouse

Virtual Warehouse

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

Data Mart

Data mart contains a subset of organization-wide data. This subset of data is valuable to specific groups of an organization.

In other words, we can claim that data marts contain data specific to a particular group. For example, the marketing data mart may contain data related to items, customers, and sales. Data marts are confined to subjects.

Points to remember about data marts −

- Window-based or Unix/Linux-based servers are used to implement data marts. They are implemented on low-cost servers.

- The implementation data mart cycles is measured in short periods of time, i.e., in weeks rather than months or years.

- The life cycle of a data mart may be complex in long run, if its planning and design are not organization-wide.

- Data marts are small in size.

- Data marts are customized by department.

- The source of a data mart is departmentally structured data warehouse.

- Data mart are flexible.

-

Enterprise Warehouse

-

An enterprise warehouse collects all the information and the subjects spanning an entire organization

-
-

It provides us enterprise-wide data integration.

-
-

The data is integrated from operational systems and external information providers.

-
-

This information can vary from a few gigabytes to hundreds of gigabytes, terabytes or beyond.

-

Load Manager

This component performs the operations required to extract and load process.

The size and complexity of the load manager varies between specific solutions from one data warehouse to other.

Load Manager Architecture

The load manager performs the following functions −

-

Extract the data from source system.

-

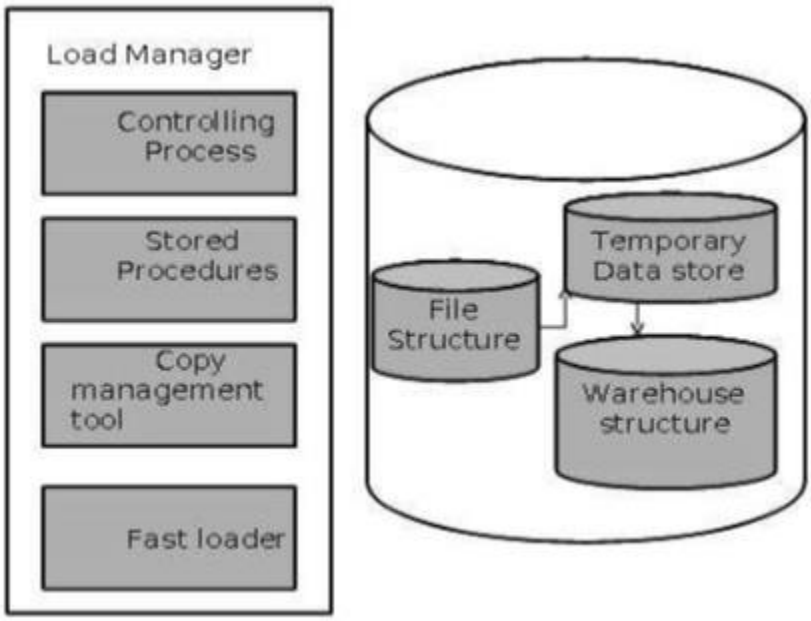
-

Fast Load the extracted data into temporary data store.

-
-

Perform simple transformations into structure similar to the one in the data warehouse.

-



Extract Data from Source

The data is extracted from the operational databases or the external information providers. Gateways is the application programs that are used to extract data. It is supported by underlying DBMS and allows client program to generate SQL to be executed at a server. Open Database Connection(ODBC), Java Database Connection (JDBC), are examples of gateway.

Fast Load

-

In order to minimize the total load window the data need to be loaded into the warehouse in the fastest possible time.

-
-

The transformations affects the speed of data processing.

-
-

It is more effective to load the data into relational database prior to applying transformations and checks.

-
-

Gateway technology proves to be not suitable, since they tend not be performant when large data volumes are involved.

-

Simple Transformations

While loading it may be required to perform simple transformations. After this has been completed we are in position to do the complex checks. Suppose we are loading the EPOS sales transaction we need to perform the following checks:

- Strip out all the columns that are not required within the warehouse.
- Convert all the values to required data types.

Warehouse Manager

A warehouse manager is responsible for the warehouse management process. It consists of third-party system software, C programs, and shell scripts.

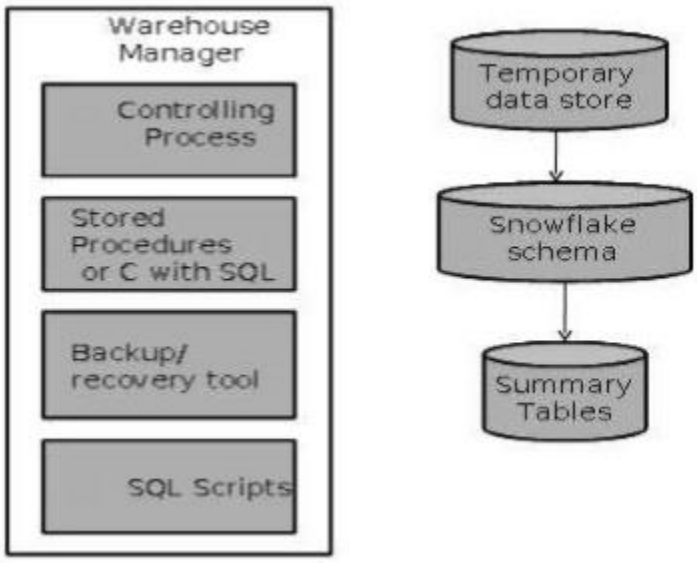
The size and complexity of warehouse managers varies between specific solutions.

Warehouse Manager Architecture

A warehouse manager includes the following −

- The controlling process
- Stored procedures or C with SQL

- Backup/Recovery tool
- SQL Scripts



Operations Performed by Warehouse Manager

-

A warehouse manager analyzes the data to perform consistency and referential integrity checks.

-
-

Creates indexes, business views, partition views against the base data.

-
-

Generates new aggregations and updates existing aggregations. Generates normalizations.

-
-

Transforms and merges the source data into the published data warehouse.

-

- Backup the data in the data warehouse.
-
- Archives the data that has reached the end of its captured life.
-

Note − A warehouse Manager also analyzes query profiles to determine index and aggregations are appropriate.

Query Manager

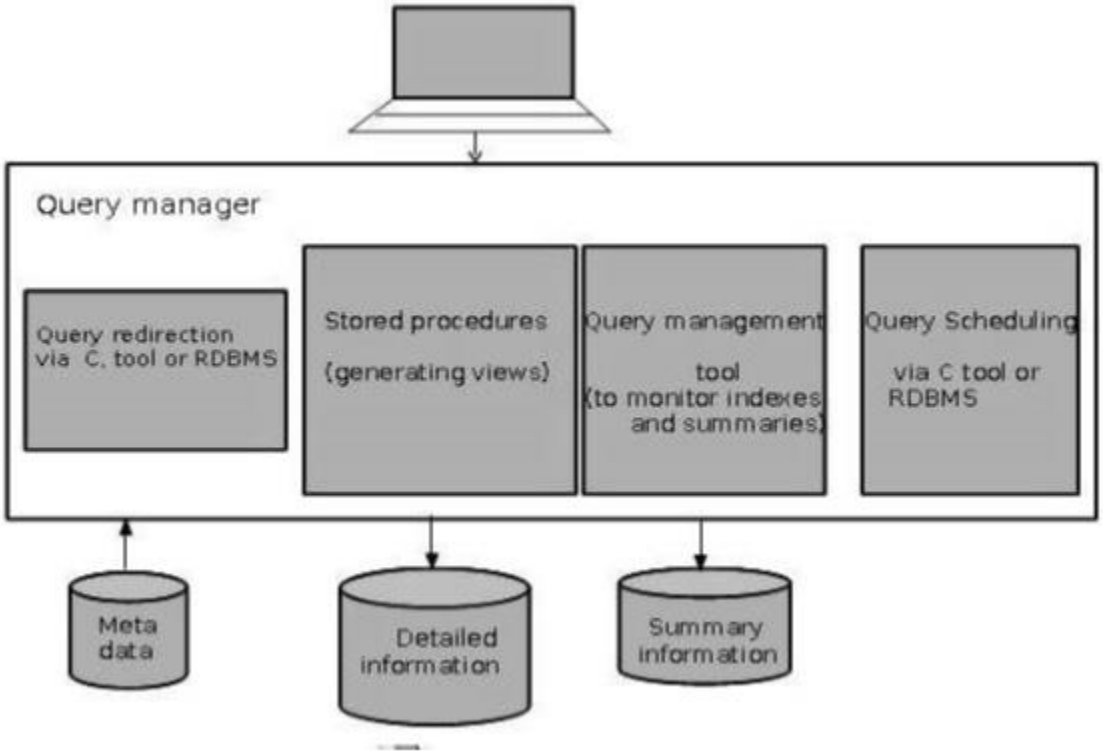
- Query manager is responsible for directing the queries to the suitable tables.
-
- By directing the queries to appropriate tables, the speed of querying and response generation can be increased.
-
- Query manager is responsible for scheduling the execution of the queries posed by the user.
-

Query Manager Architecture

The following screenshot shows the architecture of a query manager. It includes the following:

- Query redirection via C tool or RDBMS
- Stored procedures
- Query management tool

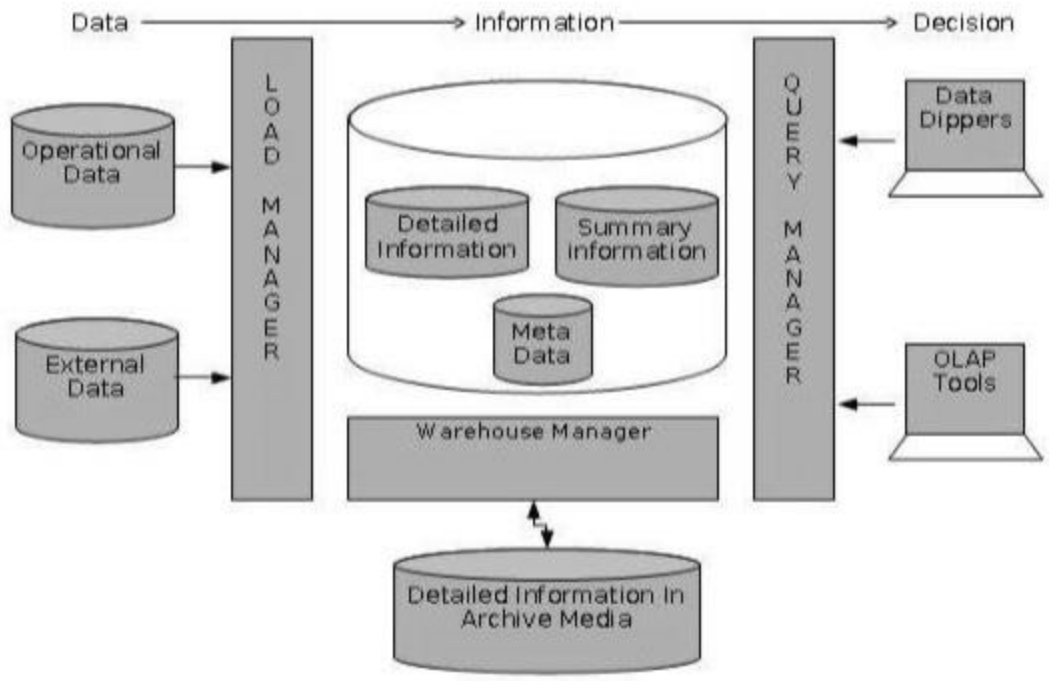
- Query scheduling via C tool or RDBMS
- Query scheduling via third-party software



Detailed Information

Detailed information is not kept online, rather it is aggregated to the next level of detail and then archived to tape. The detailed information part of data warehouse keeps the detailed information in the starflake schema. Detailed information is loaded into the data warehouse to supplement the aggregated data.

The following diagram shows a pictorial impression of where detailed information is stored and how it is used.



Unit--V

OLAP

Online Analytical Processing Server (OLAP) is based on the multidimensional data model. It allows managers, and analysts to get an insight of the information through fast, consistent, and interactive access to information. This chapter cover the types of OLAP, operations on OLAP, difference between OLAP, and statistical databases and OLTP.

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

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 −

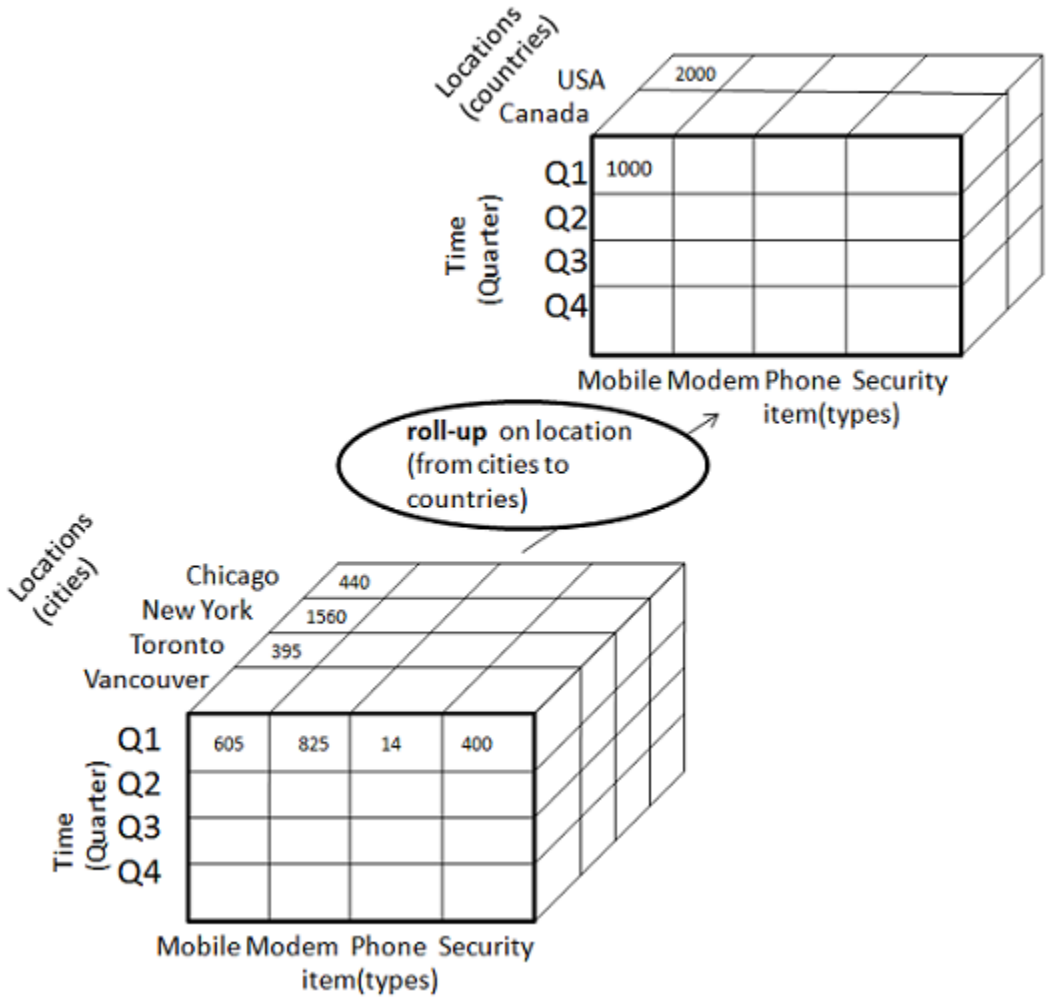
- Roll-up
- Drill-down
- Slice and dice
- Pivot (rotate)

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.

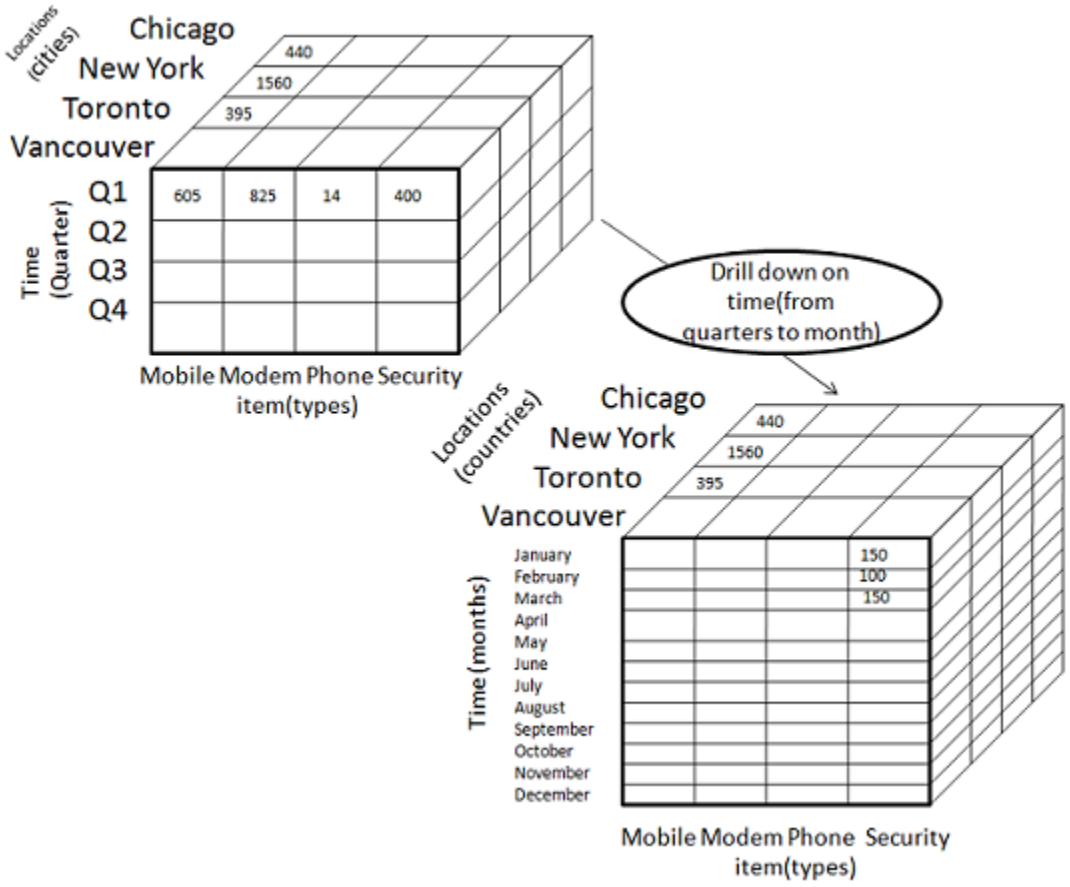


Drill-down

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

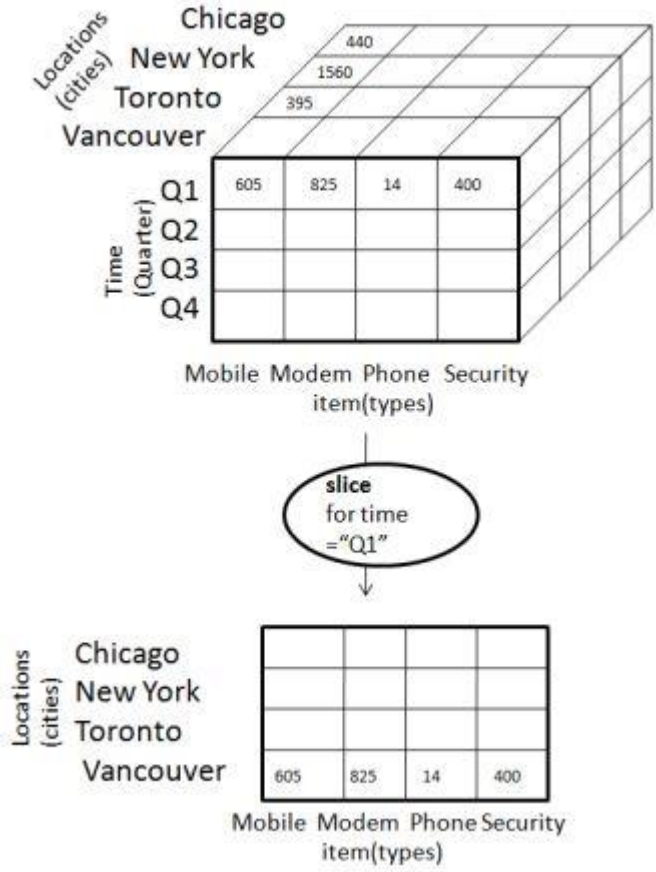
- By stepping down a concept hierarchy for a dimension
- By introducing a new dimension.

The following diagram illustrates how drill-down works −



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.



-

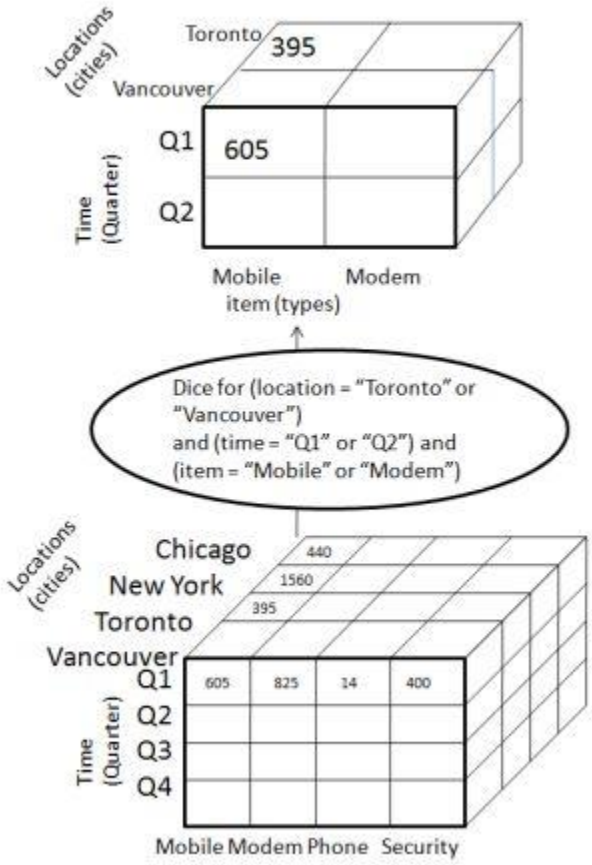
Here Slice is performed for the dimension "time" using the criterion time = "Q1".

-
-

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.

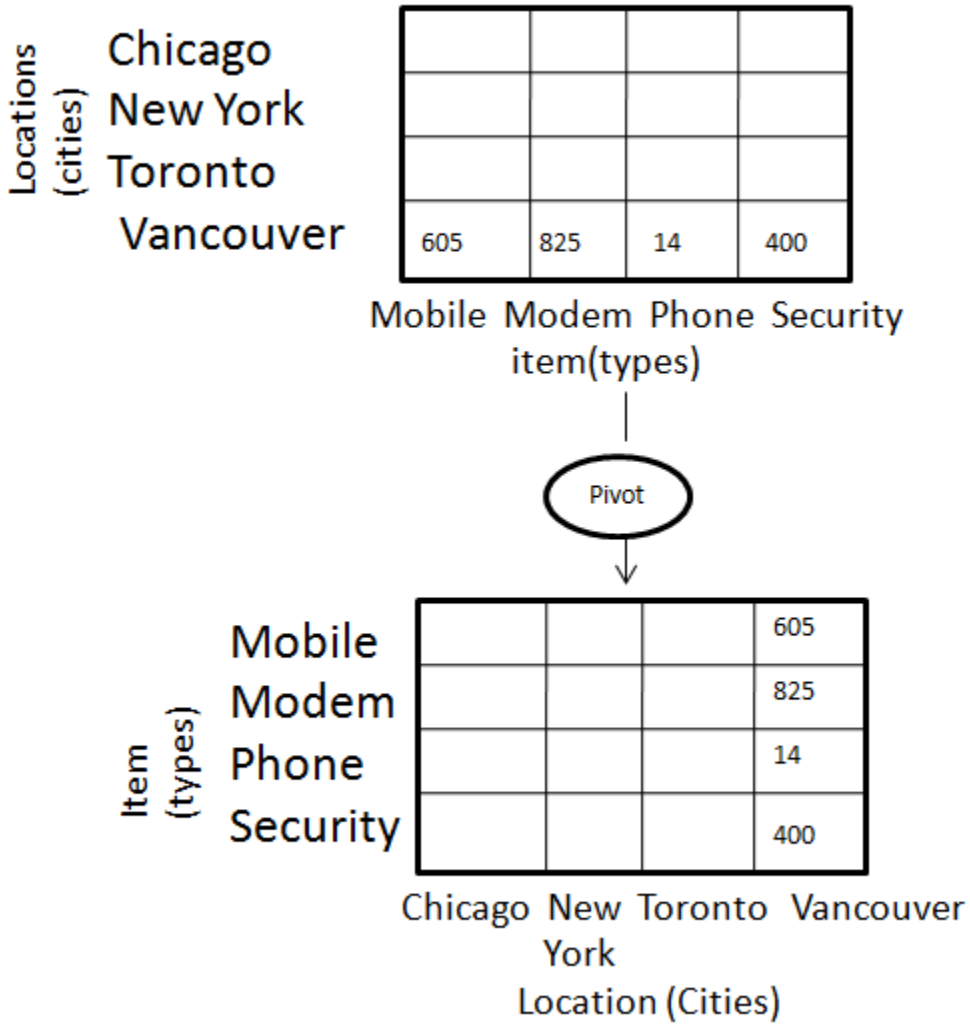


The dice operation on the cube based on the following selection criteria involves three dimensions.

- (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.