VIVA

A Data Warehousing (DW):

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 various sources.

"A data warehouse is an electronic storage of an organization's historical data for the purpose of reporting, analysis and data mining or knowledge discovery."

Characteristic

- Subject oriented
- Integrated
- Time -variant
- Non volatile

Advantages of Data Warehouse (DWH)

- To alleviate the burden on the manufacturing system, the Data Warehouse helps to combine multiple data sources.
- The data warehouse helps to decrease the overall research and reporting turnaround time.
- Restructuring and convergence make documentation and review simpler for the customer.
- The data warehouse allows users to access confidential data from a single location from several sources. It also saves time for users to access data from various sources.
- A large amount of historical information is stored in a data center. This helps users to compare different times and trends to construct possible predictions.

Disadvantages of Data Warehouse (DWH)

- Data centers are high-quality maintenance systems. The data warehouse could be impacted by any reorganization of the business processes and the source systems, resulting in high maintenance costs.
- The data warehouse may sound basic, but it's just too complex for average people.
- Despite the best intentions of project management, the scope of the data storage project will begin to grow.
- At this point, warehouse customers can have various business rules in place.

Benefit DW:

- Scalability
- Save time
- Analysis, trend, predication
- Improves the decision-making process
- Provides competitive advantage

Metadata:

Metadata is simply defined as data about data. The data that is used to represent other data is known as metadata. For example, the index of a book serves as a metadata for the contents in the book.

Role of metadata:

- Metadata acts as a directory.
- This directory helps the decision support system to locate the contents of the data warehouse.
- Metadata helps in decision support system for mapping of data when data is transformed from operational environment to data warehouse environment.
- Metadata helps in summarization between current detailed data and highly summarized data.
- Metadata also helps in summarization between lightly detailed data and highly summarized data.
- Metadata is used for query tools.
- Metadata is used in extraction and cleansing tools.
- Metadata is used in reporting tools.
- Metadata is used in transformation tools.
- Metadata plays an important role in loading functions.

A **Data Mart** is a subset of (DW)a directorial information store, generally oriented to a specific purpose or primary data subject which may be distributed to provide business needs.

S.NO	Data Warehouse	Data Mart
1.	Data warehouse is a Centralised system.	While it is a decentralised system.
2.	In data warehouse, lightly denormalization takes place.	While in Data mart, highly denormalization takes place.
3.	Data warehouse is top-down model.	While it is a bottom-up model.
4.	To built a warehouse is difficult.	While to build a mart is easy.
5.	In data warehouse, Fact constellation schema is used.	While in this, Star schema and snowflake schema are used.
6.	Data Warehouse is flexible.	While it is not flexible.
7.	Data Warehouse is the data-oriented in nature.	While it is the project-oriented in nature.
8.	Data Ware house has long life.	While data-mart has short life than warehouse.
9.	In Data Warehouse, Data are contained in detail form.	While in this, data are contained in summarized form.
10.	Data Warehouse is vast in size.	While data mart is smaller than warehouse.
11.	It collects data from various data sources.	It generally stores data from a data warehouse.
12.	Long time for processing the data because of large data.	Less time for processing the data because of handling only a small amount of data.
13.	Complicated design process of creating schemas and views.	Easy design

S.N o	ER Modeling	Dimensional Modeling
1	It is transaction-oriented.	It is subject-oriented.
2	Entities and Relationships.	Fact Tables and Dimension Tables.
3	Few levels of granularity.	Multiple levels of granularity.
4	Real-time information.	Historical information.
5	It eliminates redundancy.	It plans for redundancy.
6	High transaction volumes using few records at a time.	Low transaction volumes using many records at a time.
7	Highly Volatile data.	Non-volatile data.
8	Physical and Logical Model.	Physical Model.
9	Normalization is suggested.	De-Normalization is suggested.

S.No	ER Modeling	Dimensional Modeling
10	OLTP Application.	OLAP Application.
Ex	The application is used for buying products from e-commerce websites like Amazon.	Application to analyze buying patterns of the customer of the various cities over the past 10 years.

Fact

It is something measurable

A Fact Table contains

1. Measurements/facts
2. Foreign key to dimension table

dimension table :

- A **dimension table** contains dimensions of a fact.
- They are joined to fact table via a foreign key.

fact less fact table

fact less fact table is means only the key available in the fact there is no measure available .

Star Schema: In a star schema, the fact table will be at the center and is connected to the dimension tables.

- Adv 1. The tables are completely in a denormalized structure.
- Simplest DW schema
- Easy to understand.

Snowflake Schema : A snowflake schema is an extension of star schema where the dimension tables are connected to one or more dimensions.

- The performance of SQL queries is a bit less when compared to star schema as more number of joins are involved.
- Hybrid
- Data redundancy is low .

S.NO	Star Schema	Snowflake Schema
1.	In star schema, The fact tables and the dimension tables are contained.	While in snowflake schema, The fact tables, dimension tables as well as sub dimension tables are contained.
2.	Star schema is a top-down model.	While it is a bottom-up model.
3.	Star schema uses more space.	While it uses less space.
4.	It takes less time for the execution of queries.	While it takes more time than star schema for the execution of queries.
5.	In star schema, Normalization is not used.	While in this, Both normalization and denormalization are used.
6.	It's design is very simple.	While it's design is complex.
7.	The query complexity of star schema is low.	While the query complexity of snowflake schema is higher than star schema.
8.	It's understanding is very simple.	While it's understanding is difficult.
9.	It has less number of foreign keys.	While it has more number of foreign keys.
10.	It has high data redundancy.	While it has low data redundancy.

ETL is a process in Data Warehousing

ETL is a process in Data Warehousing and it stands for **Extract, Transform and Load**. It is a process in which an ETL tool **extracts** the data from various data source systems, **transforms** it in the staging area, and then finally, **loads** it into the Data Warehouse system.

Extraction:

The first step of the ETL process is extraction. In this step, data from various source systems is extracted which can be in various formats like relational databases, No SQL, XML, and flat files into the staging area.

Transformation:

The second step of the ETL process is transformation. In this step, a set of rules or functions are applied on the extracted data to convert it into a single standard format.

Loading:

The third and final step of the ETL process is loading. In this step, the transformed data is finally loaded into the data warehouse.

OLAP stands for On-Line 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.

Types of OLAP Servers

We have four types of OLAP servers −

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

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 reductio

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 reductio

Slice

A **slice** is a subset of the cubes corresponding to a single value for one or more members of the dimension.

Dice

The dice operation describes a sub cube by operating a selection on two or more dimension.

Pivot

The pivot operation is also called a rotation. Pivot is a visualization operations which rotates the data axes in view to provide an alternative presentation of the data.

Adv olap

- OLTP offers accurate forecast for revenue and expense.
- It provides a solid foundation for a stable business /organization due to timely modification of all transactions.
- OLTP makes transactions much easier on behalf of the customers.
- It broadens the client base for an organization by speeding up and simplifying individual processes.
- OLTP provides support for bigger databases.
- Partition of data for data manipulation is easy.
- We need OLTP to use the tasks which are frequently performed by the system.
- When we need only a small number of records.

Dis olap

- If the OLTP system faces hardware failures, then online transactions get severely affected.
- OLTP systems allow multiple users to access and change the same data at the same time, which many times created an unprecedented situation.
- If the server hangs for seconds, it can affect to a large number of transactions.
- OLTP required a lot of staff working in groups in order to maintain inventory.
- Online Transaction Processing Systems do not have proper methods of transferring products to buyers by themselves.

Application OLTP:

- Financial
- Sales
- Business

2 Introduction to Data Mining, Data Exploration and Data Pre-processing

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.**

KDD: The term KDD stands for Knowledge Discovery in Databases. It refers to the broad procedure of discovering knowledge in data and emphasizes the high-level applications of specific Data Mining techniques.

1. ***Data Cleaning***: Data cleaning is defined as removal of noisy and irrelevant data from collection.
    - Cleaning in case of ***Missing values***.
    - Cleaning ***noisy*** data, where noise is a random or variance error.
    - Cleaning with ***Data discrepancy detection*** and ***Data transformation tools***.
2. ***Data Integration***: Data integration is defined as heterogeneous data from multiple sources combined in a common source(DataWarehouse).
    - Data integration using ***Data Migration tools***.
    - Data integration using ***Data Synchronization tools***.
    - Data integration using ***ETL***(Extract-Load-Transformation) process.
3. ***Data Selection***: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
    - Data selection using ***Neural network***.
    - Data selection using ***Decision Trees***.
    - Data selection using ***Naive bayes***.
    - Data selection using ***Clustering***, ***Regression***, etc.
4. ***Data Transformation***: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure. Data Transformation is a two step process:
    - ***Data Mapping***: Assigning elements from source base to destination to capture transformations.
    - ***Code generation***: Creation of the actual transformation program.
5. ***Data Mining***: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
    - Transforms task relevant data into ***patterns***.
    - Decides purpose of model using ***classification*** or ***characterization***.
6. ***Pattern Evaluation***: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures.
    - Find ***interestingness score*** of each pattern.
    - Uses ***summarization*** and ***Visualization*** to make data understandable by user.

7. ***Knowledge representation***: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.
    - Generate ***reports***.
    - Generate ***tables***.
    - Generate ***discriminant rules***, ***classification rules***, ***characterization rules***, etc.

Data Visualization refers to the visual representation of data with the help of comprehensive charts, images, lists, charts, and other visual objects.

Data reduction techniques ensure the integrity of data while reducing the data. Data reduction is a process that reduces the volume of original data and represents it in a much smaller volume. Data reduction techniques are used to obtain a reduced representation of the dataset that is much smaller in volume by maintaining the integrity of the original data. By reducing the data, the efficiency of the data mining process is improved, which produces the same analytical results.

Data Mining Applications

Here is the list of areas where data mining is widely used −

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications
- Intrusion Detection

- 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.

Ratio Scaled attribute:

Both differences and ratios are significant for Ratio. For eg. age, length, and Weight.

3 Classification + 4

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

What is a Decision Tree Algorithm?

A decision tree is a tree in which every node is either a leaf node or a decision node. This tree takes an input an object and outputs some decision. All Paths from root node to the leaf node are reached by either using AND or OR or BOTH. The tree is constructed using the regularities of the data. The decision tree is not affected by Automatic Data Preparation.

What Is Naive Bayes Algorithm?

Naive Bayes Algorithm is used to generate mining models. These models help to identify relationships between input columns and the predictable columns. This algorithm can be used in the initial stage of exploration. The algorithm calculates the probability of every state of each input column given predictable columns possible states. After the model is made, the results can be used for exploration and making predictions.

Clustering Algorithm?

Clustering algorithm is used to group sets of data with similar characteristics also called as clusters. These clusters help in making faster decisions, and exploring data. The algorithm first identifies relationships in a dataset following which it generates a series of clusters based on the relationships. The process of creating clusters is iterative. The algorithm redefines the groupings to create clusters that better represent the data**.**

agglomerative clustering

In agglomerative clustering, each data point act as an individual cluster and at each step, data objects are grouped in a bottom-up method. Initially, each data object is in its cluster. At each iteration, the clusters are combined with different clusters until one cluster is formed.

Divisive Hierarchical Clustering

Divisive hierarchical clustering is exactly the opposite of Agglomerative Hierarchical clustering. In Divisive Hierarchical clustering, all the data points are considered an individual cluster, and in every iteration, the data points that are not similar are separated from the cluster.

Advantages of Hierarchical clustering

- It is simple to implement and gives the best output in some cases.
- It is easy and results in a hierarchy, a structure that contains more information.
- It does not need us to pre-specify the number of clusters

Disadvantages of hierarchical clustering

- It breaks the large clusters.
- It is Difficult to handle different sized clusters and convex shapes.
- It is sensitive to noise and outliers.
- The algorithm can never be changed or deleted once it was done previously

5) Mining frequent patterns and associations

Market basket analysis:

Market basket analysis is a data mining technique used by retailers to increase sales by better understanding customer purchasing patterns. It involves Analyzing large data sets, such as purchase history, to reveal product groupings and products that are likely to be purchased together.

frequent item set:

An itemset consists of two or more items. An itemset that occurs frequently is called a frequent itemset. Thus frequent itemset mining is a data mining technique to identify the items that often occur together. For Example, Bread and butter, Laptop and Antivirus software, etc

Association rule

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction. A typical example is a Market Based Analysis.

Support is an indication of how frequently the items appear in the data.

Confidence indicates the number of times the if-then statements are found true.

Apriori algorithm

Apriori algorithm refers to an algorithm that is used in mining frequent products sets and relevant association rules.

- It used to solve frequent itemset problem.
- It user bottom -up approach
- Apriori algorithm helps the customers to buy their products with ease and increases the sales performance of the particular store.

The given three components comprise the Apriori algorithm.

- Support
- Confidence
- Lift

Advantages of Apriori Algorithm

- It is used to calculate large itemset.
- Simple to understand and apply.

Disadvantages of Apriori Algorithms

- Apriori algorithm is an expensive method to find support since the calculation has to pass through the whole database.
- Sometimes, you need a huge number of candidate rules, so it becomes computationally more expensive.

Mining Frequent Itemset without candidate generation i.e. FP TREE

FP TREE :

- frequent patterns
- The FP-Growth Algorithm is an alternative way to find frequent item sets without using candidate generations, thus improving performance.
- FP TREE Is tree like structure which consist of one root node called null and sub tree
- Sub consists of 1item name 2count 3node-link

Disadvantages of FP-Growth Algorithm

This algorithm also has some disadvantages, such as:

- FP Tree is more cumbersome and difficult to build than Apriori.
- It may be expensive.
- The algorithm may not fit in the shared memory when the database is large.

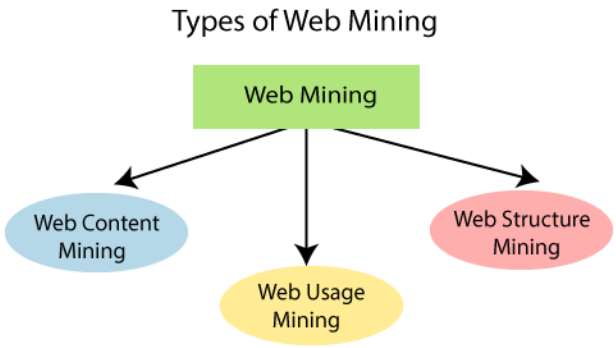
Difference between Apriori and FP Growth Algorithm

- Apriori and FP-Growth algorithms are the most basic FIM algorithms. There are some basic differences between these algorithms

Web mining

Web Mining is the process of Data Mining techniques to automatically discover and extract information from Web documents and services. The main purpose of web mining is discovering useful information from the World-Wide Web and its usage patterns.

There are three types of data mining:



Web Content Mining: Web content mining is the application of extracting useful information from the content of the web documents. Web content consist of several types of data – text, image, audio, video etc. Content data is the group of facts that a web page is designed.

Web Structure Mining: Web structure mining is the application of discovering structure information from the web. The structure of the web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages.

Web Usage Mining: Web usage mining is the application of identifying or discovering interesting usage patterns from large data sets. And these patterns enable you to understand the user behaviours or something like that.

Application of Web Mining:

Web mining has an extensive application because of various uses of the web. The list of some applications of web mining is given below.

- Marketing and conversion tool
- Data analysis on website and application accomplishment.
- Audience behavior analysis
- Advertising and campaign accomplishment analysis.
- Testing and analysis of a site.

Web structure mining technique:

- Page rank
- Clever

PageRank (PR) is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages. According to Google:

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

Extra

Mention Some Of The Data Mining Techniques?

- ○ Statistics
- ○ Machine learning ○
- Decision Tree
- ○ Hidden markov models
- ○ Artificial Intelligence
- ○ Genetic Algorithm ○ Meta learning