What is a Data Warehouse?

A Data Warehouse (DW) is a relational database that is designed for query and analysis rather than transaction processing. It includes historical data derived from transaction data from single and multiple sources.

A Data Warehouse provides integrated, enterprise-wide, historical data and focuses on providing support for decision-makers for data modeling and analysis.

A Data Warehouse is a group of data specific to the entire organization, not only to a particular group of users.

It is not used for daily operations and transaction processing but used for making decisions.

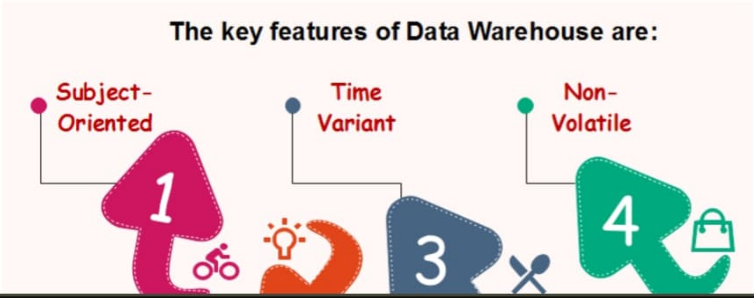
It is not used for daily operations and transaction processing but used for making decisions.

A Data Warehouse can be viewed as a data system with the following attributes:

- It is a database designed for investigative tasks, using data from various applications.
- It supports a relatively small number of clients with relatively long interactions.
- It includes current and historical data to provide a historical perspective of information.
- Its usage is read-intensive.
- It contains a few large tables.

"Data Warehouse is a subject-oriented, integrated, and time-variant store of information in support of management's decisions."

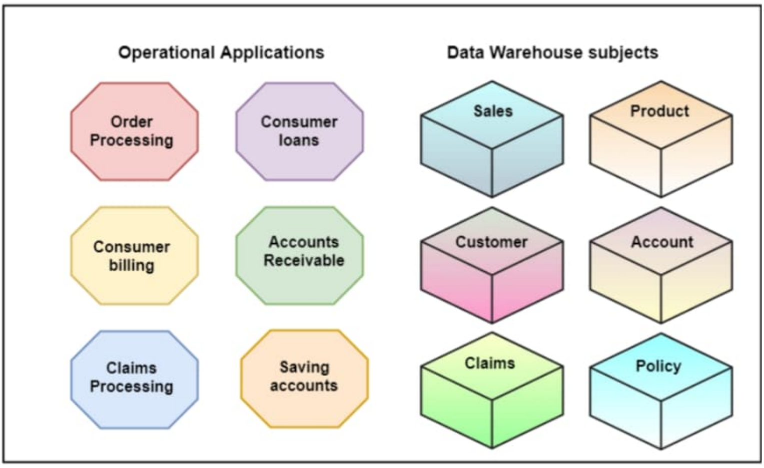
Characteristics of Data Warehouse



Subject-Oriented

A data warehouse target on the modeling and analysis of data for decision-makers. Therefore, data warehouses typically provide a concise and straightforward view around a particular subject, such as customer, product, or sales, instead of the global organization's ongoing operations. This is done by excluding data that are not useful concerning the subject and including all data needed by the users to understand the subject.

Data Warehouse is Subject-Oriented



Integrated

A data warehouse integrates various heterogeneous data sources like RDBMS, flat files, and online transaction records. It requires performing data cleaning and integration during data warehousing to ensure consistency in naming conventions, attributes types, etc.,

Time-Variant

Historical information is kept in a data warehouse. For example, one can retrieve files from 3 months, 6 months, 12 months, or even previous data from a data warehouse. These variations with a transactions system, where often only the most current file is kept.

Non-Volatile

The data warehouse is a physically separate data storage, which is transformed from the source operational RDBMS. The operational updates of data do not occur in the data warehouse, i.e., update, insert, and delete operations are not performed. It usually requires only two procedures in data accessing: Initial loading of data and access to data. Therefore, the DW does not require transaction processing, recovery, and concurrency capabilities, which allows for substantial speedup of data retrieval. Non-Volatile defines that once entered into the

Difference between Database System and Data Warehouse

Database System: Database System is used in traditional way of storing and retrieving data. The major task of database system is to perform query processing. These systems are generally referred as online transaction processing system. These systems are used day to day operations of any organization. Data Warehouse: Data Warehouse is the place where huge amount of data is stored. It is meant for users or knowledge workers in the role of data analysis and decision making. These systems are supposed to organize and present data in different format and different forms in order to serve the need of the specific user for specific purpose. These systems are referred as online analytical processing. **Difference between Database System and Data Warehouse:**

Database System	Data Warehouse
It supports operational processes.	It supports analysis and performance reporting.
Capture and maintain the data.	Explore the data.
Current data.	Multiple years of history.
Data is balanced within the scope of this one system.	Data must be integrated and balanced from multiple system.
Data is updated when transaction occurs.	Data is updated on scheduled processes.
Data verification occurs when entry is done.	Data verification occurs after the fact.
100 MB to GB.	100 GB to TB.
ER based.	Star/Snowflake.
Application oriented.	Subject oriented.
Primitive and highly detailed.	Summarized and consolidated.
Flat relational.	Multidimensional.

Last Updated : 09 Nov, 2022

8:10

google.com/searc

MEREL BARENDS

Differences Between Databases and Data Warehouses

Database	Data Warehouse
An organized collection of data.	A central repository of integrated data from one or more sources.
Usually tied to a single application such as a ticketing system	Usually store data from any number of applications
Primarily insert/write data	Primarily read/retrive data
Data is normalized to allow quick response times.	Data is denormalized for analytical and reporting efficiencies.
Current/Point-in-time data	Historical data
Online Transactional Processing	Online Analytical Processing
Provides a detailed relational view	Provides a summarized multidimensional view
For many concurrent transactions	Not for a large amount of concurrent transactions

difference between database and data warehouse, databas...

Visit >

$9.90*

... - www.merelbarends.com

Share Save

of storing data about every sale of clothing, we can store total sales of clothing by item name and category.

Need for Data Warehouse

An ordinary Database can store MBs to GBs of data and that too for a specific purpose. For storing data of TB size, the storage shifted to the Data Warehouse. Besides this, a transactional database doesn't offer itself to analytics. To effectively perform analytics, an organization keeps a central Data Warehouse to closely study its business by organizing, understanding, and using its historical data for making strategic decisions and analyzing trends.

Benefits of Data Warehouse

- **Better business analytics:** Data warehouse plays an important role in every business to store and analysis of all the past data and records of the company. which can further increase the understanding or analysis of data for the company.
- **Faster Queries: The data** warehouse is designed to handle large queries that's why it runs queries faster than the database.
- **Improved data Quality:** In the data warehouse the data you gathered from different sources is being stored and analyzed it does not interfere with or add data by itself so your quality of data is maintained and if you get any issue regarding data quality then the data warehouse team will solve this.
- **Historical Insight:** The warehouse stores all your historical data which contains details about the business so that one can analyze it at any time and extract insights from it.

Data Warehouse vs DBMS

Database	Data Warehouse
A common Database is based on operational or transactional processing. Each operation is an indivisible transaction.	A data Warehouse is based on analytical processing.
Generally, a Database stores current and up-to-date data which is used for daily operations.	A Data Warehouse maintains historical data over time. Historical data is the data kept over years and can used for trend analysis, make future predictions and decision support.
A database is generally application specific. **Example** – A database stores related data, such as the student details in a school.	A Data Warehouse is integrated generally at the organization level, by combining data from different databases. **Example** – A data warehouse integrates the data from one or more databases , so that analysis can be done to get results , such as the best performing school in a city.
Constructing a Database is not so expensive.	Constructing a Data Warehouse can be expensive.

Example Applications of Data Warehousing

Data Warehousing can be applied anywhere where we have a huge amount of data and we want to see statistical results that help in decision making.

Need for Data Warehouse

Data Warehouse is needed for the following reasons:

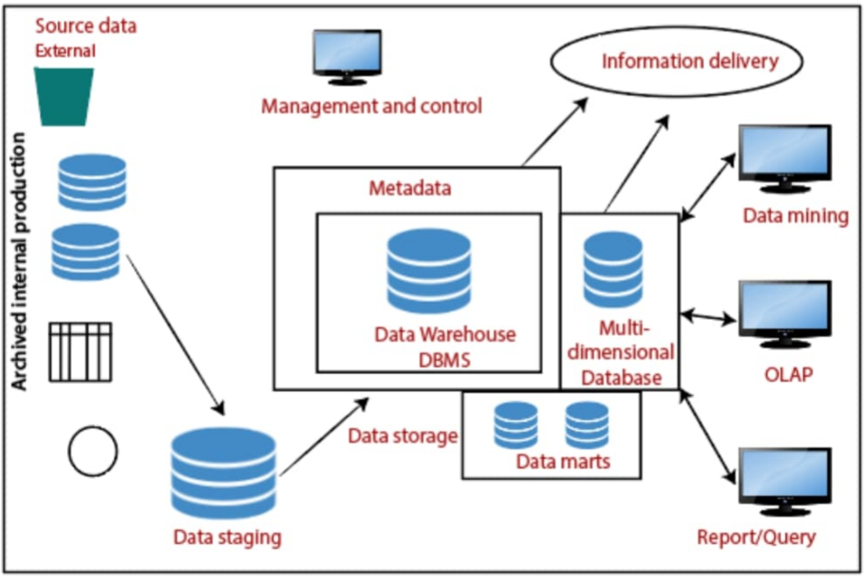
1. 1) **Business User:** Business users require a data warehouse to view summarized data from the past. Since these people are non-technical, the data may be presented to them in an elementary form.
2. 2) **Store historical data:** Data Warehouse is required to store the time variable data from the past. This input is made to be used for various purposes.
3. 3) **Make strategic decisions:** Some strategies may be depending upon the data in the data warehouse. So, data warehouse contributes to making strategic decisions.
4. 4) **For data consistency and quality:** Bringing the data from different sources at a commonplace, the user can effectively undertake to bring the uniformity and consistency in data.
5. 5) **High response time:** Data warehouse has to be ready for somewhat unexpected loads and types of queries, which demands a significant degree of flexibility and quick response time.

Benefits of Data Warehouse

1. Understand business trends and make better

Components or Building Blocks of Data Warehouse

Architecture is the proper arrangement of the elements. We build a data warehouse with software and hardware components. To suit the requirements of our organizations, we arrange these building we may want to boost up another part with extra tools and services. All of these depends on our circumstances.



Components or Building Blocks of Data Warehouse

The figure shows the essential elements of a

Source Data Component

Source data coming into the data warehouses may be grouped into four broad categories:

Production Data: This type of data comes from the different operating systems of the enterprise. Based on the data requirements in the data warehouse, we choose segments of the data from the various operational modes.

Internal Data: In each organization, the client keeps their "**private**" spreadsheets, reports, customer profiles, and sometimes even department databases. This is the internal data, part of which could be useful in a data warehouse.

Archived Data: Operational systems are mainly intended to run the current business. In every operational system, we periodically take the old data and store it in achieved files.

External Data: Most executives depend on

External Data: Most executives depend on information from external sources for a large percentage of the information they use. They use statistics associating to their industry produced by the external department.

Data Staging Component

After we have been extracted data from various operational systems and external sources, we have to prepare the files for storing in the data warehouse. The extracted data coming from several different sources need to be changed, converted, and made ready in a format that is relevant to be saved for querying and analysis.

We will now discuss the three primary functions that take place in the staging area.

1) Data Extraction: This method has to deal with numerous data sources. We have to employ the appropriate techniques for each data source.

1) Data Extraction: This method has to deal with numerous data sources. We have to employ the appropriate techniques for each data source.

2) Data Transformation: As we know, data for a data warehouse comes from many different sources. If data extraction for a data warehouse posture big challenges, data transformation present even significant challenges. We perform several individual tasks as part of data transformation.

First, we clean the data extracted from each source. Cleaning may be the correction of misspellings or may deal with providing default values for missing data elements, or elimination of duplicates when we bring in the same data from various source systems.

Standardization of data components forms a large part of data transformation. Data transformation contains many forms of combining pieces of data from different sources. We combine data from single source

On the other hand, data transformation also contains purging source data that is not useful and separating outsource records into new combinations. Sorting and merging of data take place on a large scale in the data staging area. When the data transformation function ends, we have a collection of integrated data that is cleaned, standardized, and summarized.

3) Data Loading: Two distinct categories of tasks form data loading functions. When we complete the structure and construction of the data warehouse and go live for the first time, we do the initial loading of the information into the data warehouse storage. The initial load moves high volumes of data using up a substantial amount of time.

Data Storage Components

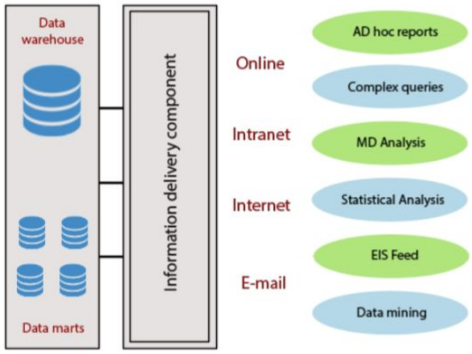
Data storage for the data warehousing is a split repository. The data repositories for the operational systems generally include only the current data. Also, these data repositories include the data structured in highly normalized for fast and efficient processing

Information Delivery Compon

The information delivery eler

Information Delivery Component

The information delivery element is used to enable the process of subscribing for data warehouse files and having it transferred to one or more destinations according to some customer-specified scheduling algorithm.



Information delivery component

Metadata Component

Metadata in a data warehouse is equal to the data dictionary or the data catalog in a database management system. In the data dictionary, we keep the data about the logical data structures, the data about the records and addresses, the information about the indexes, and so on.

Data Marts

Data Marts

It includes a subset of corporate-wide data that is of value to a specific group of users. The scope is confined to particular selected subjects. Data in a data warehouse should be a fairly current, but not mainly up to the minute, although development in the data warehouse industry has made standard and incremental data dumps more achievable. Data marts are lower than data warehouses and usually contain organization. The current trends in data warehousing are to developed a data warehouse with several smaller related data marts for particular kinds of queries and reports.

Management and Control Component

The management and control elements coordinate the services and functions within the data warehouse. These components control the data transformation and the data transfer into the data warehouse storage. On the other hand, it moderates the data delivery to the clients. Its work with the database management systems and au

be correctly saved in the

monitors the movement of info

Data Warehouse Modeling

Data warehouse modeling is the process of designing the schemas of the detailed and summarized information of the data warehouse. The goal of data warehouse modeling is to develop a schema describing the reality, or at least a part of the fact, which the data warehouse is needed to support.

Data warehouse modeling is an essential stage of building a data warehouse for two main reasons. Firstly, through the schema, data warehouse clients can visualize the relationships among the warehouse data, to use them with greater ease. Secondly, a well-designed schema allows an effective data warehouse structure to emerge, to help decrease the cost of implementing the warehouse and improve the efficiency of using it.

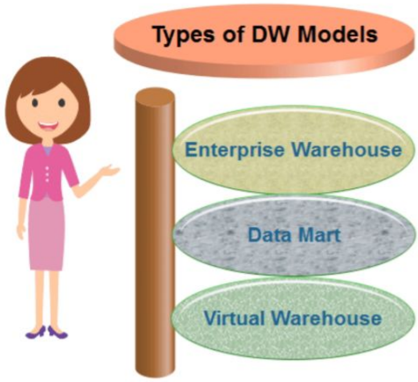
Data modeling in data warehouses is different from data modeling in operational database systems. The primary function of data warehouses is to support DSS processes. Thus, the objective of data warehouse modeling is to make the data warehouse

Data modeling in data warehouses is different from data modeling in operational database systems. The primary function of data warehouses is to support DSS processes. Thus, the objective of data warehouse modeling is to make the data warehouse efficiently support complex queries on long term information.

In contrast, data modeling in operational database systems targets efficiently supporting simple transactions in the database such as retrieving, inserting, deleting, and changing data. Moreover, data warehouses are designed for the customer with general information knowledge about the enterprise, whereas operational database systems are more oriented toward use by software specialists for creating distinct applications.

Data Warehouse model is illustrated in the given diagram

Types of Data Warehouse Models



Enterprise Warehouse

An Enterprise warehouse collects all of the records about subjects spanning the entire organization. It supports corporate-wide data integration, usually from one or more operational systems or external data providers, and it's cross-functional in scope. It generally contains detailed information as well as summarized information and can range in estimate from a few gigabyte to hundreds of gigabytes, terabytes, or beyond.

An enterprise data warehouse may be accomplished on traditional mainframes, UNIX super servers, or parallel architecture platforms. It required extensive business estimate from a few gigabyte to hundreds of gigabytes, terabytes, or beyond.

An enterprise data warehouse may be accomplished on traditional mainframes, UNIX super servers, or parallel architecture platforms. It required extensive business modeling and may take years to develop and build.

Data Mart

A data mart includes a subset of corporate-wide data that is of value to a specific collection of users. The scope is confined to particular selected subjects. For example, a marketing data mart may restrict its subjects to the customer, items, and sales. The data contained in the data marts tend to be summarized.

Data Marts is divided into two parts:

Independent Data Mart: Independent data mart is sourced from data captured from one or more operational systems or external data providers, or data generally locally within a different department or geographic area.

Dependent Data Mart: Dependent data marts are sourced exactly from enterprise data-is sourced from data captured from one or more operational systems or external data providers, or data generally locally within a different department or geographic area.

Dependent Data Mart: Dependent data marts are sourced exactly from enterprise data-warehouses.

Virtual Warehouses

Virtual Data Warehouses is a set of perception over the operational database. For effective query processing, only some of the possible summary vision may be materialized. A virtual warehouse is simple to build but required excess capacity on operational database servers.

← Prev

Next →

What is Meta Data in Data Warehousing?

Metadata is data that describes and contextualizes other data. It provides information about the content, format, structure, and other characteristics of data, and can be used to improve the organization, discoverability, and accessibility of data.

Metadata can be stored in various forms, such as text, XML, or RDF, and can be organized using metadata standards and schemas. There are many metadata standards that have been developed to facilitate the creation and management of metadata, such as Dublin Core, schema.org, and the Metadata Encoding and Transmission Standard (METS). Metadata schemas define the structure and format of metadata and provide a consistent framework for organizing and describing data.

Several Examples of Metadata:

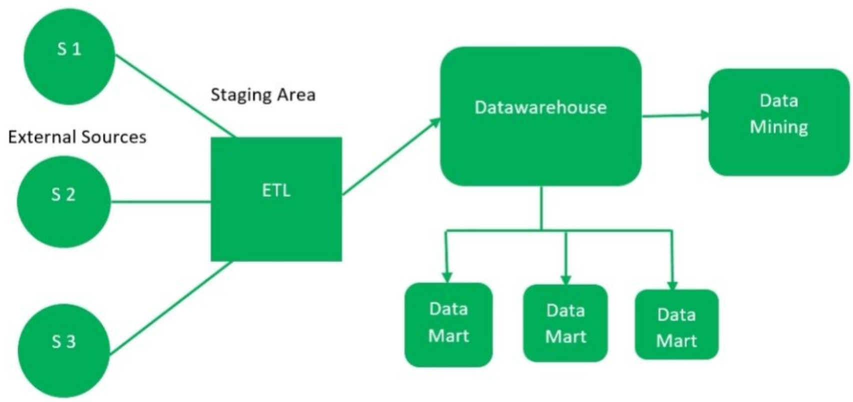
Metadata is data that provides information about other data. Here are a few examples of metadata:

1. **File metadata:** This includes information about a file, such as its name, size, type, and creation date.
2. **Image metadata:** This includes information about an image, such as its resolution, color depth, and camera settings.
3. **Music metadata:** This includes information about a piece of music, such as its title, artist, album, and genre.
4. **Video metadata:** This includes information about a video, such as its length, resolution, and frame rate.
5. **Document metadata:** This includes information about a document, such as its author, title, and creation date.
6. **Database metadata:** This includes information about a database, such as its structure, tables, and fields.
7. **Web metadata:** This includes information about a we h as its title,

Data Warehouse Architecture

A **data-warehouse** is a heterogeneous collection of different data sources organised under a unified schema. There are 2 approaches for constructing data-warehouse: Top-down approach and Bottom-up approach are explained as below.

1. Top-down approach:



The essential components are discussed

1. **External Sources –**
   External source is a source from where data is collected irrespective of the type of data. Data can be structured, semi structured and unstructured as well.

2. **Stage Area –**
   Since the data, extracted from the external sources does not follow a particular format, so there is a need to validate this data to load into datawarehouse. For this purpose, it is recommended to use **ETL** tool.
   - **E(Extracted):** Data is extracted from External data source.
   - **T(Transform):** Data is transformed into the standard format.
   - **L(Load):** Data is loaded into datawarehouse after transforming it into the standard format.

3. **Data-warehouse –**
   After cleansing of data, it is stored in the datawarehouse as central repository. It actually stores the meta data and the actual data gets stored in the data marts. **Note** that datawarehouse stores the data in its purest form in this top-down approach.

4. **Data Marts –**
   Data mart is also a part of storage component. It stores the information of a particular function of an organisation which is handled by single authority. There can be as many number of data marts in an organisation depending upon the functions. We can also say that data mart contains subset of the data stored in datawarehouse.

5. **Data Mining –**
   The practice of analysing the big data present in datawarehouse is data mining. It is used to find the hidden patterns that are present in the database or in datawarehouse with the help of algorithm of data mining.
   This approach is defined by **Inmon** as – datawarehouse as a central repository for the complete organisation and data marts are created from it after the complete datawarehouse has been created.

Advantages of Top-Down Approach –

1. Since the data marts are created from the datawarehouse, provides consistent dimensional view of data marts.

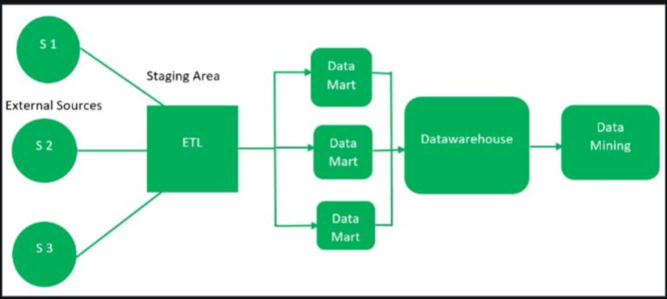
2. Also, this model is considered as the strongest model for business changes. That's why, big organisations prefer to follow this approach.

3. Creating data mart from datawarehouse is easy.
4. Improved data consistency: The top-down approach promotes data consistency by ensuring that all data marts are sourced from a common data warehouse. This ensures that all data is standardized, reducing the risk of errors and inconsistencies in reporting.
5. Easier maintenance: Since all data marts are sourced from a central data warehouse, it is easier to maintain and update the data in a top-down approach. Changes can be made to the data warehouse, and those changes will automatically propagate to all the data marts that rely on it.
6. Better scalability: The top-down approach is highly scalable, allowing organizations to add new data marts as needed without disrupting the existing infrastructure. This is particularly important for organizations that are experiencing rapid growth or have evolving business needs.
7. Improved governance: The top-down approach facilitates better governance by enabling centralized control of data access, security, and quality. This ensures that all data is managed consistently and that it meets the organization's standards for quality and compliance.
8. Reduced duplication: The top-down approach reduces data duplication by ensuring that data is stored only once in the data warehouse. This saves storage space and reduces the risk of data inconsistencies.
9. Better reporting: The top-down approach enables better reporting by providing a consistent view of data across all data marts. This makes it easier to create accurate and timely reports, which can improve decision-making and drive better business outcomes.
10. Better data integration: The top-down approach enables better data integration by ensuring that all data marts are sourced from a common data warehouse. This makes it easier to integrate data from different sources and provides a more complete view of the organization's data.

Disadvantages of Top-Down Approach –

1. The cost, time taken in designing and its maintenance is very high.
2. Complexity: The top-down approach can be complex to implement and maintain, particularly for large organizations with complex data needs. The design and implementation of the data warehouse and data marts can be time-consuming and costly.
3. Lack of flexibility: The top-down approach may not be suitable for organizations that require a high degree of flexibility in their data reporting and analysis. Since the design of the data warehouse and data marts is pre-determined, it may not be possible to adapt to new or changing business requirements.
4. Limited user involvement: The top-down approach can be dominated by IT departments, which may lead to limited user involvement in the design and implementation process. This can result in data marts that do not meet the specific needs of business users.
5. Data latency: The top-down approach may result in data latency, particularly when data is sourced from multiple systems. This can impact the accuracy and timeliness of reporting and analysis.
6. Data ownership: The top-down approach can create challenges around data ownership and control. Since data is centralized in the data warehouse, it may not be clear who is responsible for maintaining and updating the data.
7. Cost: The top-down approach can be expensive to implement and maintain, particularly for smaller organizations that may not have the resources to invest in a large-scale data warehouse and associated data marts.
8. Integration challenges: The top-down approach may face challenges in integrating data from different sources, particularly when data is stored in different formats or structures. This can lead to data inconsistencies and inaccuracies.

2. Bottom-up approach:



1. First, the data is extracted from external sources (same as happens in top-down approach).

2. Then, the data go through the staging area (as explained above) and loaded into data marts instead of datawarehouse. The data marts are created first and provide reporting capability. It addresses a single business area.

3. These data marts are then integrated into datawarehouse.

This approach is given by **Kinball** as – data marts are created first and provides a thin view for analyses and datawarehouse is created after complete data marts have been created.

Advantages of Bottom-Up Approach –

1. As the data marts are created first, so the reports are quickly generated.

2. We can accommodate more number of data marts here and in this way datawarehouse can be extended.

3. Also, the cost and time taken in designing this model is low comparatively.
4. Incremental development: The bottom-up approach supports incremental development, allowing for the creation of data marts one at a time. This allows for quick wins and incremental improvements in data reporting and analysis.
5. User involvement: The bottom-up approach encourages user involvement in the design and implementation process. Business users can provide feedback on the data marts and reports, helping to ensure that the data marts meet their specific needs.
6. Flexibility: The bottom-up approach is more flexible than the top-down approach, as it allows for the creation of data marts based on specific business needs. This approach can be particularly useful for organizations that require a high degree of flexibility in their reporting and analysis.
7. Faster time to value: The bottom-up approach can deliver faster time to value, as the data marts can be created more quickly than a centralized data warehouse. This can be particularly useful for smaller organizations with limited resources.
8. Reduced risk: The bottom-up approach reduces the risk of failure, as data marts can be tested and

2. Then, the data go through the staging area (as explained above) and loaded into data marts instead of datawarehouse. The data marts are created first and provide reporting capability. It addresses a single business area.

3. These data marts are then integrated into datawarehouse.

This approach is given by **Kinball** as – data marts are created first and provides a thin view for analyses and datawarehouse is created after complete data marts have been created.

Advantages of Bottom-Up Approach –

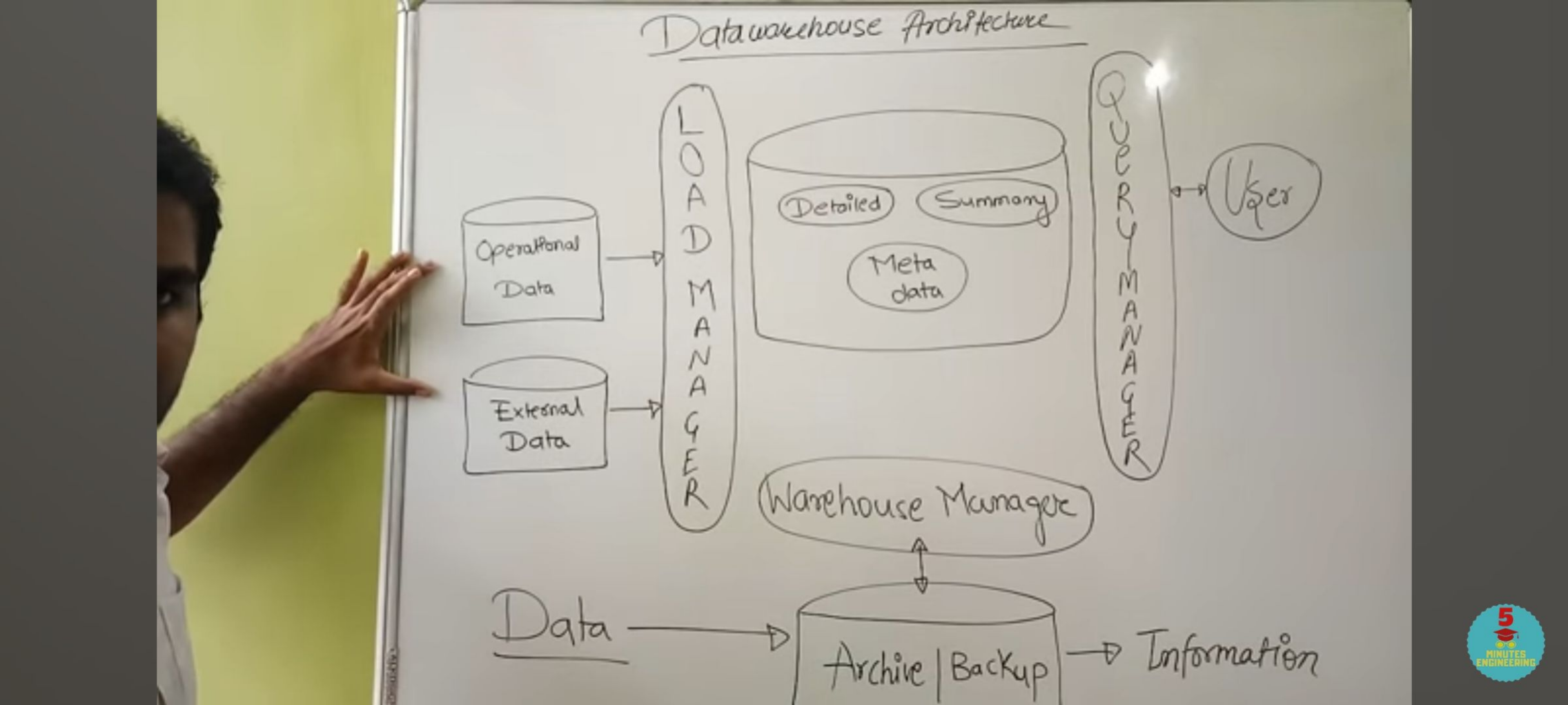
1. As the data marts are created first, so the reports are quickly generated.

2. We can accommodate more number of data marts here and in this way datawarehouse can be extended.

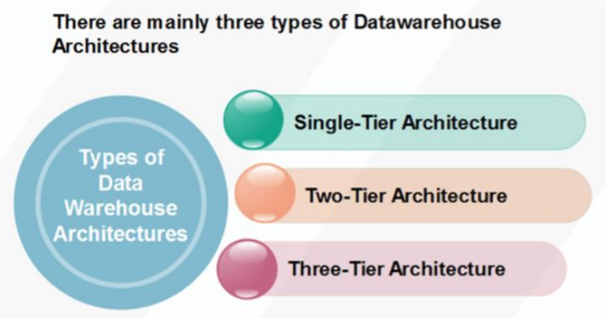
3. Also, the cost and time taken in designing this model is low comparatively.
4. Incremental development: The bottom-up approach supports incremental development, allowing for the creation of data marts one at a time. This allows for quick wins and incremental improvements in data reporting and analysis.
5. User involvement: The bottom-up approach encourages user involvement in the design and implementation process. Business users can provide feedback on the data marts and reports, helping to ensure that the data marts meet their specific needs.
6. Flexibility: The bottom-up approach is more flexible than the top-down approach, as it allows for the creation of data marts based on specific business needs. This approach can be particularly useful for organizations that require a high degree of flexibility in their reporting and analysis.
7. Faster time to value: The bottom-up approach can deliver faster time to value, as the data marts can be created more quickly than a centralized data warehouse. This can be particularly useful for smaller organizations with limited resources.
8. Reduced risk: The bottom-up approach reduces the risk of failure, as data marts can be tested and refined before being incorporated into a larger data warehouse. This approach can also help to identify and address potential data quality issues early in the process.
9. Scalability: The bottom-up approach can be scaled up over time, as new data marts can be added as needed. This approach can be particularly useful for organizations that are growing rapidly or undergoing significant change.
10. Data ownership: The bottom-up approach can help to clarify data ownership and control, as each data mart is typically owned and managed by a specific business unit. This can help to ensure that data is accurate and up-to-date, and that it is being used in a consistent and appropriate way across the organization.

Disadvantage of Bottom-Up Approach –

1. This model is not strong as top-down approach as dimensional view of data marts is not consistent as it is in above approach.
2. Data silos: The bottom-up approach can lead to the creation of data silos, where different business units create their own data marts without considering the needs of other parts of the organization. This can lead to inconsistencies and redundancies in the data, as well as difficulties in integrating data across the organization.
3. Integration challenges: Because the bottom-up approach relies on the integration of multiple data marts, it can be more difficult to integrate data from different sources and ensure consistency across the organization. This can lead to issues with data quality and accuracy.
4. Duplication of effort: In a bottom-up approach, different business units may duplicate effort by creating their own data marts with similar or overlapping data. This can lead to inefficiencies and higher costs in data management.
5. Lack of enterprise-wide view: The bottom-up approach can result in a lack of enterprise-wide view, as data marts are typically designed to meet the needs of specific business units rather than the organization as a whole. This can make it difficult to gain a comprehensive understanding of the organization's data and business processes.
6. Complexity: The bottom-up approach can be more complex than the top-down approach, as it involves the integration of multiple data marts with varying levels of complexity and granularity. This can make it more difficult to manage and maintain the data warehouse over time.
7. Risk of inconsistency: Because the bottom-up approach allows for the creation of data marts with different structures and granularities, there is a risk of inconsistency in the data. This can make it difficult to compare data across different parts of the organization or to ensure that reports are accurate and reliable.


Datawarehouse Architecture
Operational Data
External Data
LOAD MANAGER
Detailed
Summary
Meta data
QUERY MANAGER
User
Warehouse Manager
Data
Archive | Backup
Information
5 MINUTES ENGINEERING

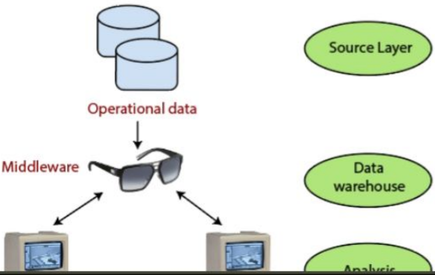
Types of Data Warehouse Architectures



Single-Tier Architecture

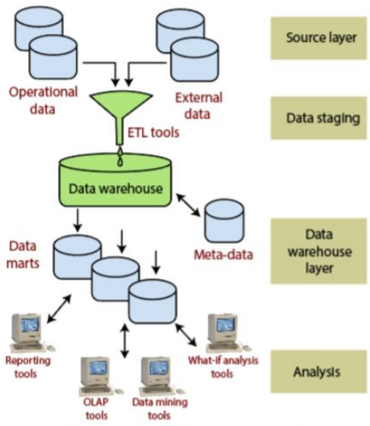
Single-Tier architecture is not periodically used in practice. Its purpose is to minimize the amount of data stored to reach this goal; it removes data redundancies.

The figure shows the only layer physically available is the source layer. In this method, data warehouses are virtual. This means that the data warehouse is implemented as a multidimensional view of operational data created by specific middleware, or an intermediate processing layer.



Two-Tier Architecture

The requirement for separation plays an essential role in defining the two-tier architecture for a data warehouse system, as shown in fig:



Two-Tier Data Warehouse Architecture

Although it is typically called two-layer architecture to highlight a separation between physically available sources and data warehouses, in fact, consists of four subsequent data flow stages:

1. **Source layer:** A data warehouse system uses a heterogeneous source of data. That data is stored initially to corporate relational databases or legacy databases, or it may come from an information system outside the corporate walls.
2. **Data Staging:** The data stored to the source should be extracted, cleansed to remove inconsistencies and fill gaps, and integrated to merge heterogeneous sources into one standard schema. The so-named **Extraction**, **Transformation**, and **Loading Tools (ETL)**

1. **Source layer:** A data warehouse system uses a heterogeneous source of data. That data is stored initially to corporate relational databases or legacy databases, or it may come from an information system outside the corporate walls.
2. **Data Staging:** The data stored to the source should be extracted, cleansed to remove inconsistencies and fill gaps, and integrated to merge heterogeneous sources into one standard schema. The so-named **Extraction**, **Transformation**, and **Loading Tools (ETL)** can combine heterogeneous schemata, extract, transform, cleanse, validate, filter, and load source data into a data warehouse.
3. **Data Warehouse layer:** Information is saved to one logically centralized individual repository: a data warehouse. The data warehouses can be directly accessed, but it can also be used as a source for creating data marts, which partially replicate data warehouse contents and are designed for specific enterprise departments. Meta-data repositories store information on sources, access procedures, data staging, users, data mart schema, and so on.
4. **Analysis:** In this layer, integrated data is efficiently, and flexible accessed to issue reports, dynamically analyze information, and simulate hypothetical business scenarios. It should feature aggregate information navigators, complex query optimizers, and customer-friendly GUIs.

Three-Tier Architecture

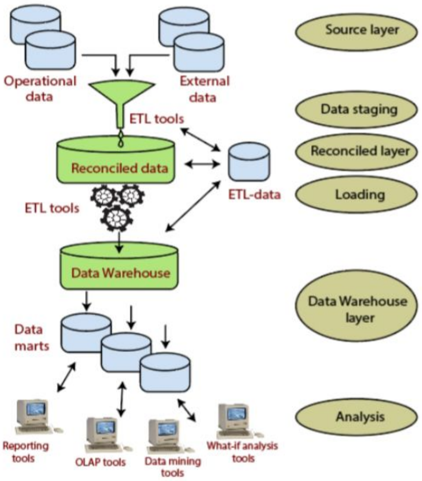
The three-tier architecture consists of the source layer (containing multiple source system), the reconciled layer and the data

Three-Tier Architecture

The three-tier architecture consists of the source layer (containing multiple source system), the reconciled layer and the data warehouse layer (containing both data warehouses and data marts). The reconciled layer sits between the source data and data warehouse.

The main advantage of the **reconciled layer** is that it creates a standard reference data model for a whole enterprise. At the same time, it separates the problems of source data extraction and integration from those of data warehouse population. In some cases, the **reconciled layer** is also directly used to accomplish better some operational tasks, such as producing daily reports that cannot be satisfactorily prepared using the corporate applications or generating data flows to feed external processes periodically to benefit from cleaning and integration.

This architecture is especially useful for the extensive, enterprise-wide systems. A disadvantage of this structure is the extra file storage space used through the extra redundant reconciled layer. It also makes the analytical tools a little further away from being real-time.



Three-Tier Architecture for a data warehouse system

D...ENSIONAL MODELLING

... in a Warehouse are usually multidimensional having ...sure and dimension attributes.

→ measure some Values and Can be aggregated upon those Values. [sum(), avg() ...]

→ that define the dimensions on which the measure attributed and their summary func^n work upon.

(BOOKSHOP)

IMP.

FACTS are the numerical measures or quantities by which one Can analyze relationships b/w dimensions.

The relation Containing such Multidimensional data are Called FACT TABLES.

Bid	tId	num...
B1	1	2...
B2	2	

Fact ...

IMP.

DIMENSIONS are the Collection of logically related attributes and is viewed as an axis for modelling the data.

A DIMENSION TABLE is a table associated with each dimension and helps in describing the dimension further.

Bid	Autho...
B1	ABC
B2	XYZ

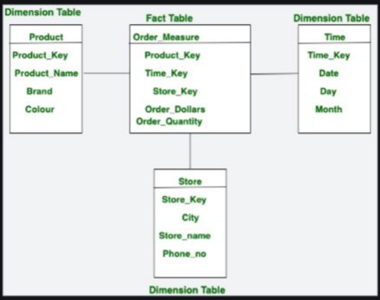
(Boo...

Difference between Fact Table and Dimension Table

A reality or fact table's record could be a combination of attributes from totally different dimension tables. The **Fact Table or Reality Table** helps the user to investigate the business dimensions that helps him in call taking to enhance his business.

On the opposite hand, **Dimension Tables** facilitate the reality table or fact table to gather dimensions on that the measures needs to be taken.

The main difference between fact table or reality table and the Dimension table is that dimension table contains attributes on that measures are taken actually table.



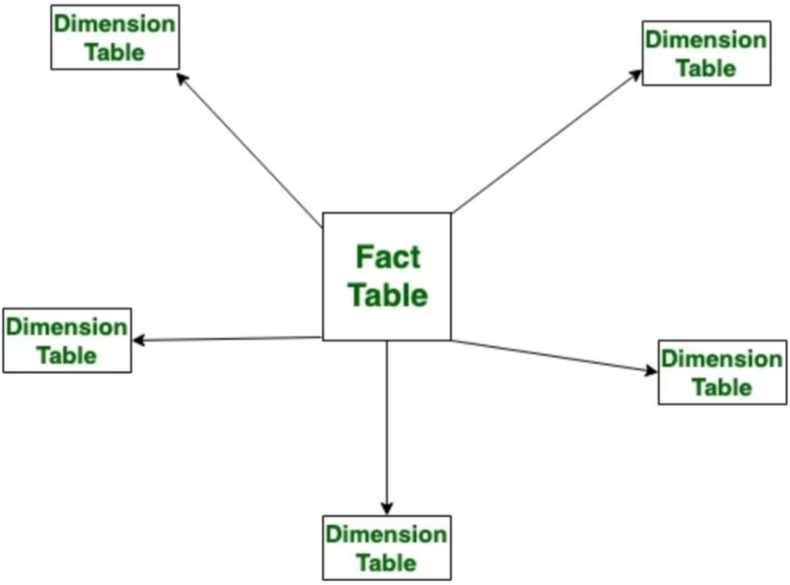
Difference between Fact Table and Dimension Table:

S.NO	Fact Table	Dimension Table
1.	Fact table contains the measuring of the attributes of a dimension table.	Dimension table contains the attributes on that truth table calculates the metric.
2.	In fact table, There is less attributes than dimension table.	While in dimension table, There is more attributes than fact table.
3.	In fact table, There is more records than dimension table.	While in dimension table, There is less records than fact table.
4.	Fact table forms a vertical table.	While dimension table forms a horizontal table.
5.	The attribute format of fact table is in numerical format and text format.	While the attribute format of dimension table is in text format.
6.	It comes after dimension table.	While it comes before fact table.
7.	The number of fact table is less than dimension table in a schema.	While the number of dimension is more than fact table in a schema.
8.	It is used for analysis purpose and decision making.	While the main task of dimension table is to store the information about a business and its process.

Also let us see what Aggregate Fact Tables are,

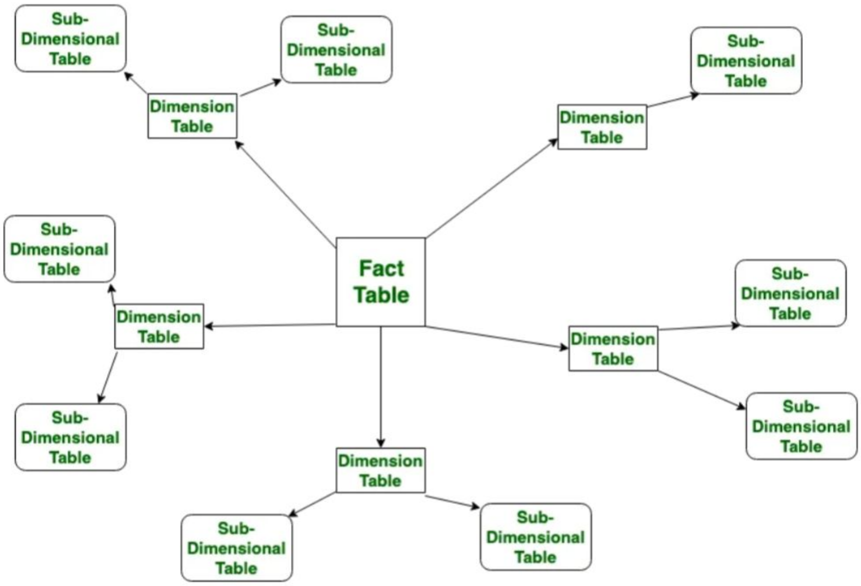
Difference between Star Schema and Snowflake Schema

Star Schema: Star schema is the type of multidimensional model which is used for data warehouse. In star schema, The fact tables and the dimension tables are contained. In this schema fewer foreign-key join is used. This schema forms a star with fact table and dimension tables.



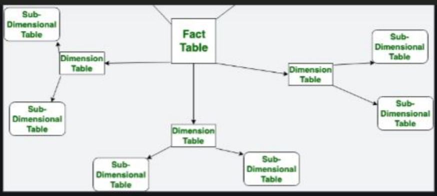
Snowflake Schema: Snowflake Schema is also the type of multidimensional model which is used for data warehouse. In snowflake schema, The fact tables, dimension tables as well as sub dimension tables are

Snowflake Schema: Snowflake Schema is also the type of multidimensional model which is used for data warehouse. In snowflake schema, The fact tables, dimension tables as well as sub dimension tables are contained. This schema forms a snowflake with fact tables, dimension tables as well as sub-dimension tables.



Let's see the difference between Star and Snowflake Schema:

S.NO	Star Schema	Snowflake Schema
		While in snowflake



Let's see the difference between Star and Snowflake Schema:

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.

Last Updated : 21 Feb, 2023

Previous
Difference between Aerospike and AllegroGraph

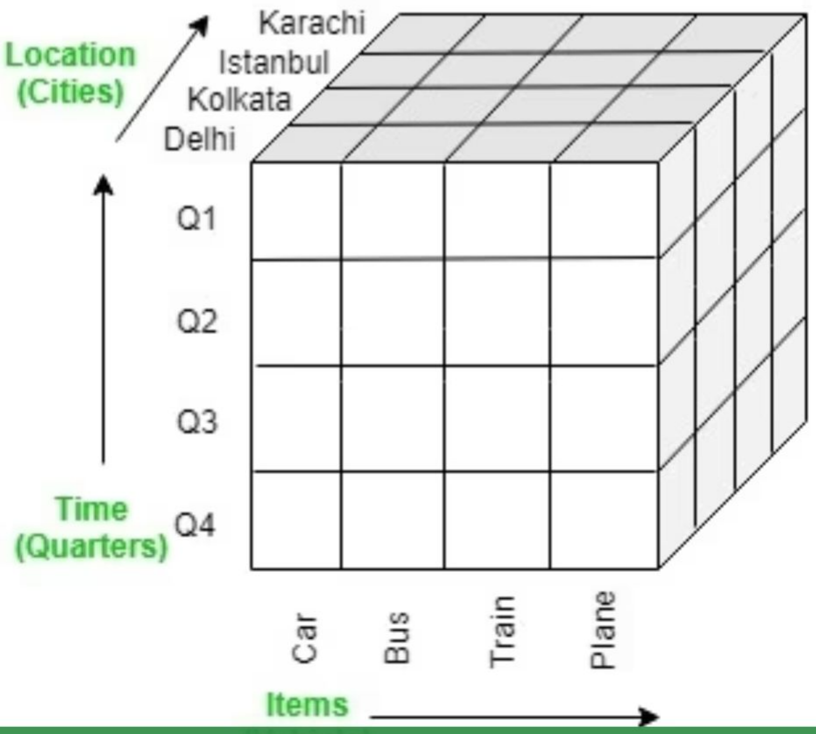
Next
Difference between Load Testing and Stress Testing

Similar Reads

Difference between Snowflake Schema and Fact Constellation Schema
Difference between Star Schema and Fact Constellation Schema

OLAP Operations in DBMS

OLAP stands for ***Online Analytical Processing*** Server. It is a software technology that allows users to analyze information from multiple database systems at the same time. It is based on multidimensional data model and allows the user to query on multi-dimensional data (eg. Delhi -> 2018 -> Sales data). OLAP databases are divided into one or more cubes and these cubes are known as *Hyper-cubes*.

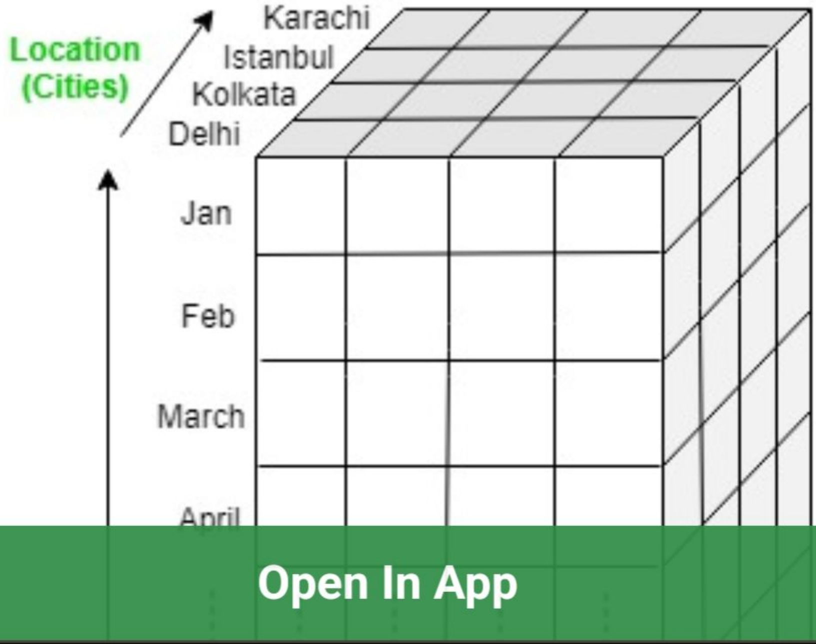


OLAP operations:

There are five basic analytical operations that can be performed on an OLAP cube:

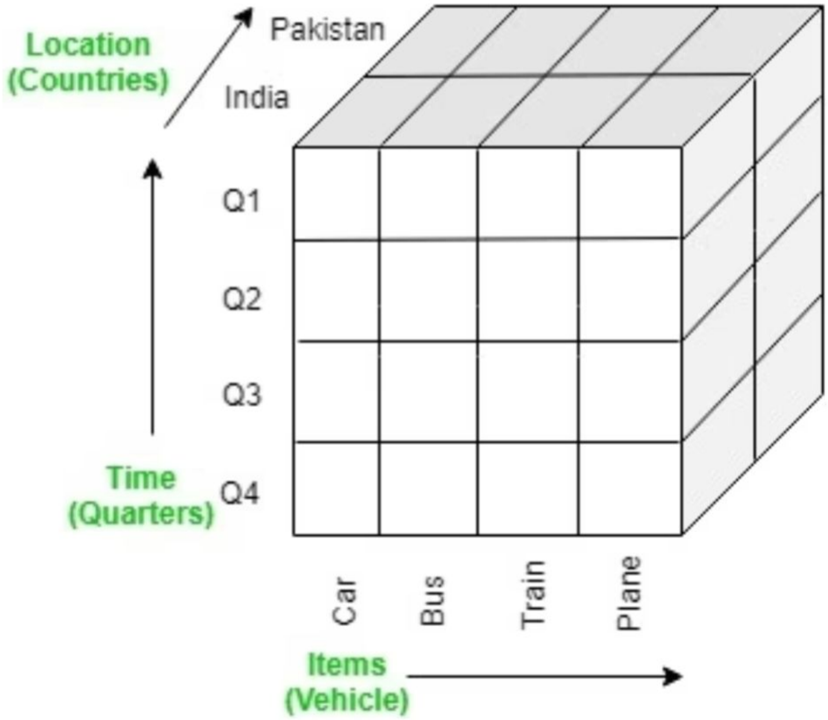
1. **Drill down:** In drill-down operation, the less detailed data is converted into highly detailed data. It can be done by:
    - Moving down in the concept hierarchy
    - Adding a new dimension

    In the cube given in overview section, the drill down operation is performed by moving down in the concept hierarchy of *Time* dimension (Quarter -> Month).



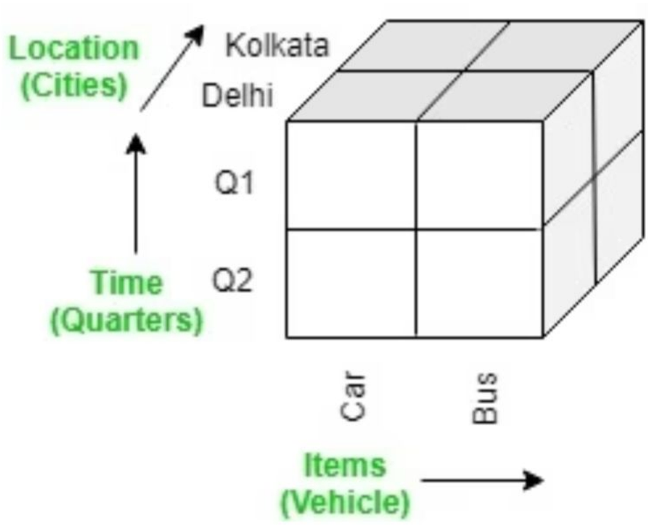
2. **Roll up:** It is just opposite of the drill-down operation. It performs aggregation on the OLAP cube. It can be done by:
   - Climbing up in the concept hierarchy
   - Reducing the dimensions

   In the cube given in the overview section, the roll-up operation is performed by climbing up in the concept hierarchy of *Location* dimension (City -> Country).



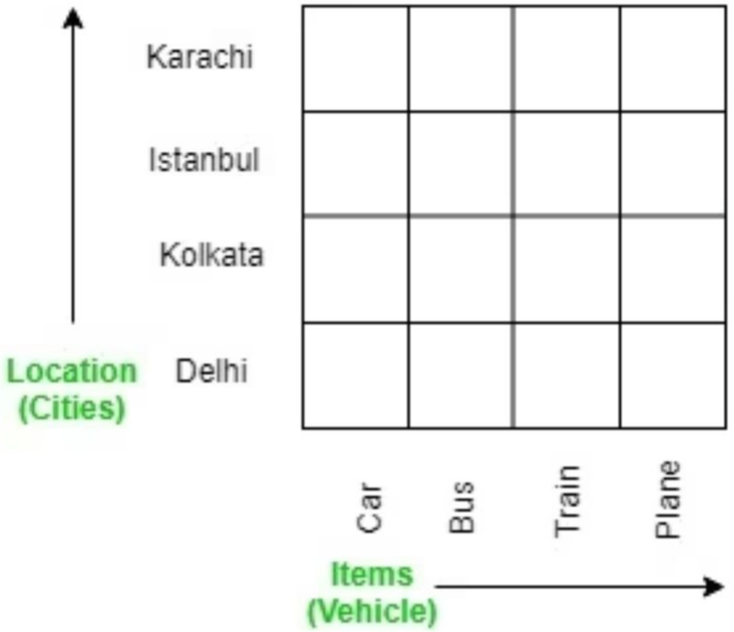
3. **Dice:** It selects a sub-cube from the OLAP cube by selec... re dimensions.

3. **Dice:** It selects a sub-cube from the OLAP cube by selecting two or more dimensions. In the cube given in the overview section, a sub-cube is selected by selecting following dimensions with criteria:
    - Location = “Delhi” or “Kolkata”
    - Time = “Q1” or “Q2”
    - Item = “Car” or “Bus”



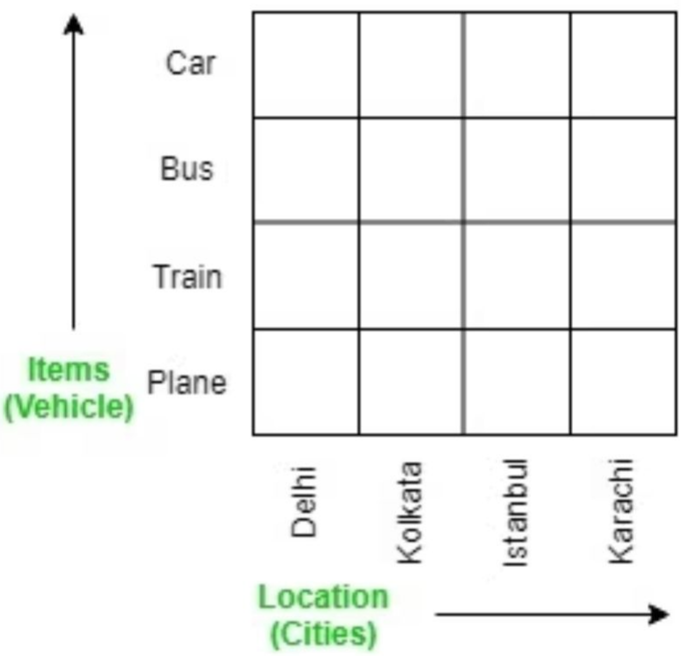
4. **Slice:** It selects a single dimension from the OLAP cube which results in a new sub-cube creation. In the cube given in the overview section, Slice is performed on the dimension Time = “Q1”.

4. **Slice:** It selects a single dimension from the OLAP cube which results in a new sub-cube creation. In the cube given in the overview section, Slice is performed on the dimension Time = “Q1”.



5. **Pivot:** It is also known as *rotation* operation as it rotates the current view to get a new view of the representation. In the sub-cube obtained after the slice operation, performing pivot operation gives a new view of it.

5. **Pivot:** It is also known as *rotation* operation as it rotates the current view to get a new view of the representation. In the sub-cube obtained after the slice operation, performing pivot operation gives a new view of it.



Article Tags : Computer Subject DBMS

Read Full Article

Difference between ROLAP, MOLAP and HOLAP

1. Relational Online Analytical Processing (ROLAP) :
ROLAP servers are placed between relational backend server and client front-end tools. It uses relational or extended DBMS to store and manage warehouse data. ROLAP has basically 3 main components: Database Server, ROLAP server, and Front-end tool.

Advantages of ROLAP –

- ROLAP is used for handle the large amount of data.
- ROLAP tools don't use pre-calculated data cubes.
- Data can be stored efficiently.
- ROLAP can leverage functionalities inherent in the relational database.

Disadvantages of ROLAP –

- Performance of ROLAP can be slow.
- In ROALP, difficult to maintain aggregate tables.
- Limited by SQL functionalities.

2. Multidimensional Online Analytical Processing (MOLAP) :
MOLAP does not uses relational database to storage.It stores in optimized multidimensional array storage. The storage utilization may be low With multidimensional data stores. Many MOLAP server handle dense and sparse data sets by using two levels of data storage representation. MOLAP has 3 components : Database Server, MOLAP server, and Front-end tool.

Advantages of MOLAP –

- MOLAP is basically used for complex calculations.
- MOLAP is optimal for operation such as slice and dice.
- MOLAP allows fastest indexing to the pre-computed summarized data.

Disadvantages of MOLAP –

- MOLAP can't handle large amount of data.
- In MOLAP, Requires additional investment.
- Without re-aggregation, difficult to change dimension.

3. Hybrid Online Analytical Processing (HOLAP) :
Hybrid is a combination of both ROLAP and MOLAP.It offers functionalities of both ROLAP and as well as MOLAP like faster computation of MOLAP and higher scalability of ROLAP. The aggregations are stored separately in MOLAP store. Its server allows storing the large data volumes of detailed information.

Advantages of HOLAP –

- HOLAP provides the functionalities of both MOLAP and ROLAP.
- HOLAP provides fast access at all levels of aggregation.

  Disadvantages of HOLAP –
  HOLAP architecture is very complex to understand because it supports both MOLAP and ROLAP.

Difference between ROLAP, MOLAP and HOLAP :

Basis	ROLAP	MOLAP	HOLAP
Storage location for summary aggregation	Relational Database is used as storage location for summary aggregation.	Multidimensional Database is used as storage location for summary aggregation.	Multidimensional Database is used as storage location for summary aggregation.
Processing time	Processing time of ROLAP is very slow.	Processing time of MOLAP is fast.	Processing time of HOLAP is fast.
Storage space requirement	Large storage space requirement in ROLAP as compare to MOLAP and	Medium storage space requirement in MOLAP as compare to ROLAP	Small storage space requirement in HOLAP as compare to MOLAP and

2. Multidimensional Online Analytical Processing (MOLAP) :
MOLAP does not uses relational database to storage.It stores in optimized multidimensional array storage. The storage utilization may be low With multidimensional data stores. Many MOLAP server handle dense and sparse data sets by using two levels of data storage representation. MOLAP has 3 components : Database Server, MOLAP server, and Front-end tool.

Advantages of MOLAP –

- MOLAP is basically used for complex calculations.
- MOLAP is optimal for operation such as slice and dice.
- MOLAP allows fastest indexing to the pre-computed summarized data.

Disadvantages of MOLAP –

- MOLAP can't handle large amount of data.
- In MOLAP, Requires additional investment.
- Without re-aggregation, difficult to change dimension.

3. Hybrid Online Analytical Processing (HOLAP) :
Hybrid is a combination of both ROLAP and MOLAP.It offers functionalities of both ROLAP and as well as MOLAP like faster computation of MOLAP and higher scalability of ROLAP. The aggregations are stored separately in MOLAP store. Its server allows storing the large data volumes of detailed information.

Advantages of HOLAP –

- HOLAP provides the functionalities of both MOLAP and ROLAP.
- HOLAP provides fast access at all levels of aggregation.

Disadvantages of HOLAP –
HOLAP architecture is very complex to understand because it supports both MOLAP and ROLAP.

Difference between ROLAP, MOLAP and HOLAP :

Basis	ROLAP	MOLAP	HOLAP
Storage location for summary aggregation	Relational Database is used as storage location for summary aggregation.	Multidimensional Database is used as storage location for summary aggregation.	Multidimensional Database is used as storage location for summary aggregation.
Processing time	Processing time of ROLAP is very slow.	Processing time of MOLAP is fast.	Processing time of HOLAP is fast.
Storage space requirement	Large storage space requirement in ROLAP as compare to MOLAP and HOLAP.	Medium storage space requirement in MOLAP as compare to ROLAP and HOLAP.	Small storage space requirement in HOLAP as compare to MOLAP and ROLAP.
Storage location for detail data	Relational database is used as storage location for detail data.	Multidimensional database is used as storage location for detail data.	Relational database is used as storage location for detail data.
Latency	Low latency in ROLAP as compare to MOLAP and HOLAP.	High latency in MOLAP as compare to ROLAP and HOLAP.	Medium latency in HOLAP as compare to MOLAP and ROLAP.
Query response time	Slow query response time in ROLAP as compare to MOLAP and HOLAP.	Fast query response time in MOLAP as compare to ROLAP and HOLAP.	Medium query response time in HOLAP as compare to MOLAP and ROLAP.

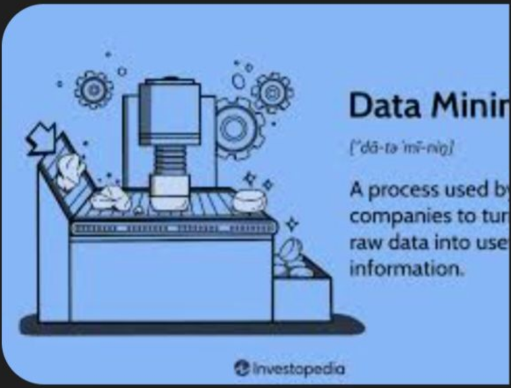
Article Tags : DBMS Difference Between data mining

डाटा माइनिंग डेफिनिशन

classification definition in data mining

Data mining

Overview | Classification | Examples



Data mining is **the process of searching and analyzing a large batch of raw data in order to identify patterns and extract useful information**. Companies use data mining software to learn more about their customers. It can help them to develop more effective marketing strategies, increase sales, and decrease costs.

https://www.investopedia.com › da...

What Is Data Mining? How It Works, Benefits, Techniques, and Examples

Feedback

People also ask

Tasks and Functionalities of Data Mining

Data Mining functions are used to define the trends or correlations contained in data mining activities. In comparison, data mining **activities** can be divided into 2 categories:

1]Descriptive Data Mining:

This category of data mining is concerned with finding patterns and relationships in the data that can provide insight into the underlying structure of the data. Descriptive data mining is often used to summarize or explore the data, and it can be used to answer questions such as: What are the most common patterns or relationships in the data? Are there any clusters or groups of data points that share common characteristics? What are the outliers in the data, and what do they represent?
Some common techniques used in descriptive data mining include:

Cluster analysis:

This technique is used to identify groups of data points that share similar characteristics. Clustering can be used for segmentation, anomaly detection, and summarization.

Association rule mining:

This technique is used to identify relationships between variables in the data. It can be used to discover co-occurring events or to identify patterns in transaction data.

Visualization:

This technique is used to represent the data in a visual format that can help users to identify patterns or trends that may not be apparent in the raw data.

2]Predictive Data Mining: This category of data mining is concerned with developing models that can predict future behavior or outcomes based on historical data. Predictive data mining is often used for classification or regression tasks, and it can be used to answer questions such as: What is the likelihood that a customer will churn? What is the expected revenue for a new product launch? What is the probability of a loan defaulting?
Some common techniques used in predictive data mining include:

Decision trees: This technique is used to create a model that can predict the value of a target variable based on the values of several input variables. Decision trees are often used for classification tasks.

Neural networks: This technique is used to create a model that can learn to recognize patterns in the data. Neural networks are often used for image recognition, speech recognition, and natural language processing.

Regression analysis: This technique is used to create a model that can predict the value of a target variable based on the values of several input variables. Regression analysis is often used for prediction tasks.

Both descriptive and predictive data mining techniques are important for gaining insights and making better decisions. Descriptive data mining can be used to explore the data and identify patterns, while predictive data mining can be used to make predictions based on those patterns. Together, these techniques can help organizations to understand their data and make informed decisions based on that understanding.

Data Mining Functionality:

1. Class/Concept Descriptions: Classes or definitions can be correlated with results. In simplified, descriptive and yet accurate ways, it can be helpful to define individual groups and concepts. These class or concept definitions are referred to as class/concept descriptions.

- **Data Characterization:** This refers to the summary of general characteristics or features of the class that is under the study. The output of the data characterization can be presented in various forms include pie charts, bar charts, curves, multidimensional data cubes.

Example: To study the characteristics of software products with sales increased by 10% in the previous years. To summarize the characteristics of the customer who spend more than $5000 a year at AllElectronics, the result is general profile of t... ers such as that they are 40-50 years old, employee and have excellent credit rating.

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 pattern
are discovered to be most c
Various types of frequency car

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 tl
set in the database.
- Confidence is the conditiona

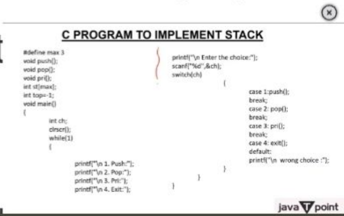
3. Association Analysis

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 collect



5. Prediction

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 busin negatively.
- ***Class predictions*** are used t

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

classes. Similar data are group

the difference being that a cl

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

grouped into any classes by t

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

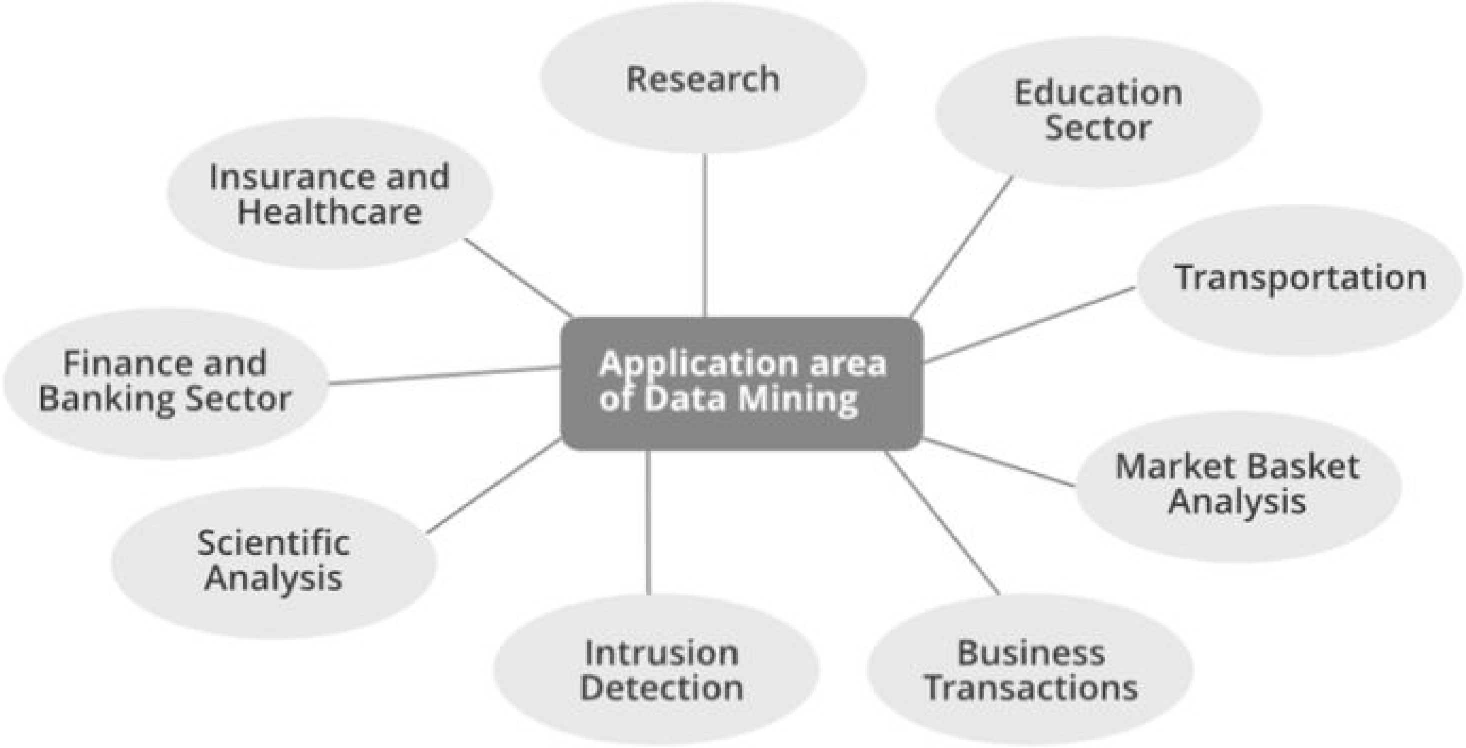
Correlation is a mathematical technique for determining whether and how strongly two attributes is related to one another. It refers to the various types of data structures, such as trees and graphs, that can be combined with an item set or subsequence. It determines how well two numerically measured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

← Prev

Applications of Data Mining

Data is a set of discrete objective facts about an event or a process that have little use by themselves unless converted into information. We have been collecting numerous data, from simple numerical measurements and text documents to more complex information such as spatial data, multimedia channels, and hypertext documents.

Nowadays, large quantities of data are being accumulated. The amount of data collected is said to be almost doubled every year. An extracting data or seeking knowledge from this massive data, data mining techniques are used. Data mining is used in almost all places where a large amount of data is stored and processed. For example, banks typically use 'data mining' to find out their prospective customers who could be interested in credit cards, personal loans, or insurance as well.

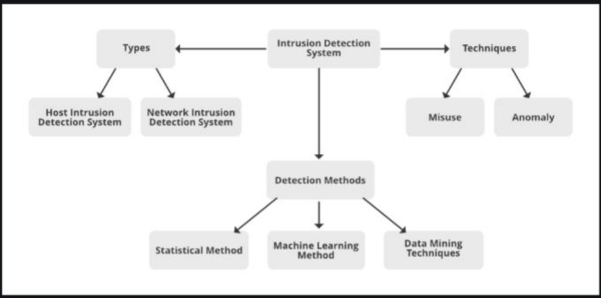

Application area of Data Mining
Research
Education Sector
Transportation
Market Basket Analysis
Business Transactions
Intrusion Detection
Scientific Analysis
Finance and Banking Sector
Insurance and Healthcare

Scientific Analysis: Scientific simulations are generating bulks of data every day. This includes data collected from nuclear laboratories, data about human psychology, etc. Data mining techniques are capable of the analysis of these data. Now we can capture and store more new data faster than we can analyze the old data already accumulated. Example of scientific analysis:

- Sequence analysis in bioinformatics
- Classification of astronomical objects
- Medical decision support.

Intrusion Detection: A network intrusion refers to any unauthorized activity on a digital network. Network intrusions often involve stealing valuable network resources. Data mining technique plays a vital role in searching intrusion detection, network attacks, and anomalies. These techniques help in selecting and refining useful and relevant information from large data sets. Data mining technique helps in classify relevant data for Intrusion Detection System. Intrusion Detection system generates alarms for the network traffic about the foreign invasions in the system. For example:

- Detect security violations
- Misuse Detection
- Anomaly Detection



Business Transactions: Every business industry is memorized for perpetuity. Such transactions are usually time-related and can be inter-business deals or intra-business operations. The effective and in-time use of the data in a reasonable time frame for competitive decision-making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world. Data mining helps to analyze these business transactions and identify marketing approaches and decision-making. Example :

- Direct mail targeting
- Stock trading
- Customer segmentation
- Churn prediction (Churn prediction is one of the most popular Big Data use cases in business)

Market Basket Analysis: Market Basket Analysis is a technique that gives the careful study of purchases done by a customer in a supermarket. This concept identifies the pattern of frequent purchase items by customers. This analysis can help to promote deals, offers, sale by the companies and data mining techniques helps to achieve this analysis task. Example:

- Data mining concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.
- Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining.
- Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc.

Education: For analyzing the education sector, data mining uses Educational Data Mining (EDM) method. This method generates patterns that can be used both by learners and educators. By using data mining EDM we can perform some educational task:

- Predicting students admission in higher education
- Predicting students profiling
- Predicting student performance
- Teachers teaching performance
- Curriculum development
- Predicting student placement opportunities

Research: A data mining technique can perform predictions, classification, clustering, associations, and grouping of data with perfection in the research area. Rules generated by data mining are unique to find results. In most of the technical research in data mining, we create a training model and testing model. The training/testing model is a strategy to measure the precision of the proposed model. It is called Train/Test because we split the data set into two sets: a training data set and a testing data set. A training data set used to design the training model whereas testing data set is used in the testing model.

- Data mining concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.
- Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining.
- Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc.

Education: For analyzing the education sector, data mining uses Educational Data Mining (EDM) method. This method generates patterns that can be used both by learners and educators. By using data mining EDM we can perform some educational task:

- Predicting students admission in higher education
- Predicting students profiling
- Predicting student performance
- Teachers teaching performance
- Curriculum development
- Predicting student placement opportunities

Research: A data mining technique can perform predictions, classification, clustering, associations, and grouping of data with perfection in the research area. Rules generated by data mining are unique to find results. In most of the technical research in data mining, we create a training model and testing model. The training/testing model is a strategy to measure the precision of the proposed model. It is called Train/Test because we split the data set into two sets: a training data set and a testing data set. A training data set used to design the training model whereas testing data set is used in the testing model. Example:

- Classification of uncertain data.
- Information-based clustering.
- Decision support system
- Web Mining
- Domain-driven data mining
- IoT (Internet of Things)and Cybersecurity
- Smart farming IoT(Internet of Things)

Healthcare and Insurance: A Pharmaceutical sector can examine its new deals force activity and their outcomes to improve the focusing of high-value physicians and figure out which promoting activities will have the best effect in the following upcoming months, Whereas the Insurance sector, data mining can help to predict which customers will buy new policies, identify behavior patterns of risky customers and identify fraudulent behavior of customers.

- Claims analysis i.e which medical procedures are claimed together.
- Identify successful medical therapies for different illnesses.
- Characterizes patient behavior to predict office visits.

Transportation: A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. A large consumer merchandise organization can apply information mining to improve its business cycle to retailers.

- Determine the distribution schedules among outlets.
- Analyze loading patterns.

Financial/Banking Sector: A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product.

- Credit card fraud detection.
- Identify 'Loyal' customers.
- Extraction of information related to customers.
- Determine credit card spending by customer groups.

Last Updated : 04 Jul, 2021

Challenges of Data Mining

Data mining, the process of extracting knowledge from data, has become increasingly important as the amount of data generated by individuals, organizations, and machines has grown exponentially. However, data mining is not without its challenges. In this article, we will explore some of the main challenges of data mining.

1]Data Quality

The quality of data used in data mining is one of the most significant challenges. The accuracy, completeness, and consistency of the data affect the accuracy of the results obtained. The data may contain errors, omissions, duplications, or inconsistencies, which may lead to inaccurate results. Moreover, the data may be incomplete, meaning that some attributes or values are

1]Data Quality

The quality of data used in data mining is one of the most significant challenges. The accuracy, completeness, and consistency of the data affect the accuracy of the results obtained. The data may contain errors, omissions, duplications, or inconsistencies, which may lead to inaccurate results. Moreover, the data may be incomplete, meaning that some attributes or values are missing, making it challenging to obtain a complete understanding of the data.

Data quality issues can arise due to a variety of reasons, including data entry errors, data storage issues, data integration problems, and data transmission errors. To address these challenges, data mining practitioners must apply data cleaning and data preprocessing techniques to improve the quality of the data. Data cleaning involves detecting and correcting errors, while data preprocessing involves transforming the data to make it suitable for data mining.

2]Data Complexity

Data complexity refers to the vast amounts of data generated by various sources, such as sensors, social media, and the internet of things (IoT). The complexity of the data may make it challenging to process, analyze, and understand. In addition, the data may be in different formats, making it challenging to integrate into a single dataset.

To address this challenge, data mining practitioners use advanced techniques such as clustering, classification, and association rule mining. These techniques help to identify patterns and relationships in the data, which can then be used to gain insights and make predictions.

3]Data Privacy and Security

Data privacy and security is another significant challenge in data mining. As more data is collected, stored, and analyzed, the risk of data breaches and cyber-attacks increases. The data may contain personal, sensitive, or con[illegible]mation that

3]Data Privacy and Security

Data privacy and security is another significant challenge in data mining. As more data is collected, stored, and analyzed, the risk of data breaches and cyber-attacks increases. The data may contain personal, sensitive, or confidential information that must be protected. Moreover, data privacy regulations such as GDPR, CCPA, and HIPAA impose strict rules on how data can be collected, used, and shared.

To address this challenge, data mining practitioners must apply data anonymization and data encryption techniques to protect the privacy and security of the data. Data anonymization involves removing personally identifiable information (PII) from the data, while data encryption involves using algorithms to encode the data to make it unreadable to unauthorized users.

4]Scalability

Data mining algorithms must be scalable to

4]Scalability

Data mining algorithms must be scalable to handle large datasets efficiently. As the size of the dataset increases, the time and computational resources required to perform data mining operations also increase. Moreover, the algorithms must be able to handle streaming data, which is generated continuously and must be processed in real-time.

To address this challenge, data mining practitioners use distributed computing frameworks such as Hadoop and Spark. These frameworks distribute the data and processing across multiple nodes, making it possible to process large datasets quickly and efficiently.

5]Interpretability

Data mining algorithms can produce complex models that are difficult to interpret. This is because the algorithms use a combination of

5]Interpretability

Data mining algorithms can produce complex models that are difficult to interpret. This is because the algorithms use a combination of statistical and mathematical techniques to identify patterns and relationships in the data. Moreover, the models may not be intuitive, making it challenging to understand how the model arrived at a particular conclusion.
To address this challenge, data mining practitioners use visualization techniques to represent the data and the models visually. Visualization makes it easier to understand the patterns and relationships in the data and to identify the most important variables.

6]Ethics

Data mining raises ethical concerns related to the collection, use, and dissemination of data. The data may be used to discriminate against certain groups, violate privacy rights, or perpetuate existing biases. Moreover, data mining algorithms may not be transparent, making it challenging to detect biases or represent the data and the models visually. Visualization makes it easier to understand the patterns and relationships in the data and to identify the most important variables.

6]Ethics

Data mining raises ethical concerns related to the collection, use, and dissemination of data. The data may be used to discriminate against certain groups, violate privacy rights, or perpetuate existing biases. Moreover, data mining algorithms may not be transparent, making it challenging to detect biases or discrimination.

KDD Process in Data Mining

Pre-requisites: Data Mining

In the context of computer science, “Data Mining” can be referred to as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Data Mining also known as Knowledge Discovery in Databases, refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases.

The need of data mining is to extract useful information from large datasets and use it to make predictions or better decision-making. Nowadays, data mining is used in almost all places where a large amount of data is stored and processed.

For examples: Banking sector, Market Basket Analysis, Network Intrusion Detection.

KDD Process

KDD (Knowledge Discovery in Databases) is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets. The KDD process is an iterative process and it requires multiple iterations of the above steps to extract accurate knowledge from the data.The following steps are included in KDD process:

Data Cleaning

Data cleaning is defined as removal of noisy and irrelevant data from collection.

1. Cleaning in case of **Missing values**.
2. Cleaning **noisy** data, where noise is a random or variance error.
3. Cleaning with **Data discrepancy detection** and **Data transformation tools**.

Data Integration

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 Synchronization tools and ETL**(Extract-Load-Transformation) process.

Data Selection

Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection. For this we can use **Neural network, Decision Trees, Naive bayes, Clustering**, and **Regression** methods.

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 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:

1. **Data Mapping**: Assigning elements from source base to destination to capture transformations.
2. **Code generation**: Creation of the actual transformation program.

Data Mining

Data mining is defined as techniques that are applied to extract patterns potentially useful. It transforms task relevant data into **patterns, and d**ecides purpose of model using **classification** or **characterization**.

Pattern Evaluation

Pattern Evaluation is defined as identifying strictly increasing patterns representing

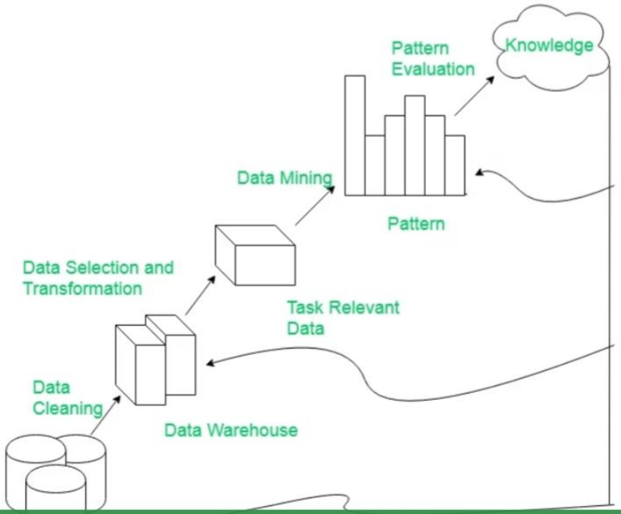
classification or **characterization**.

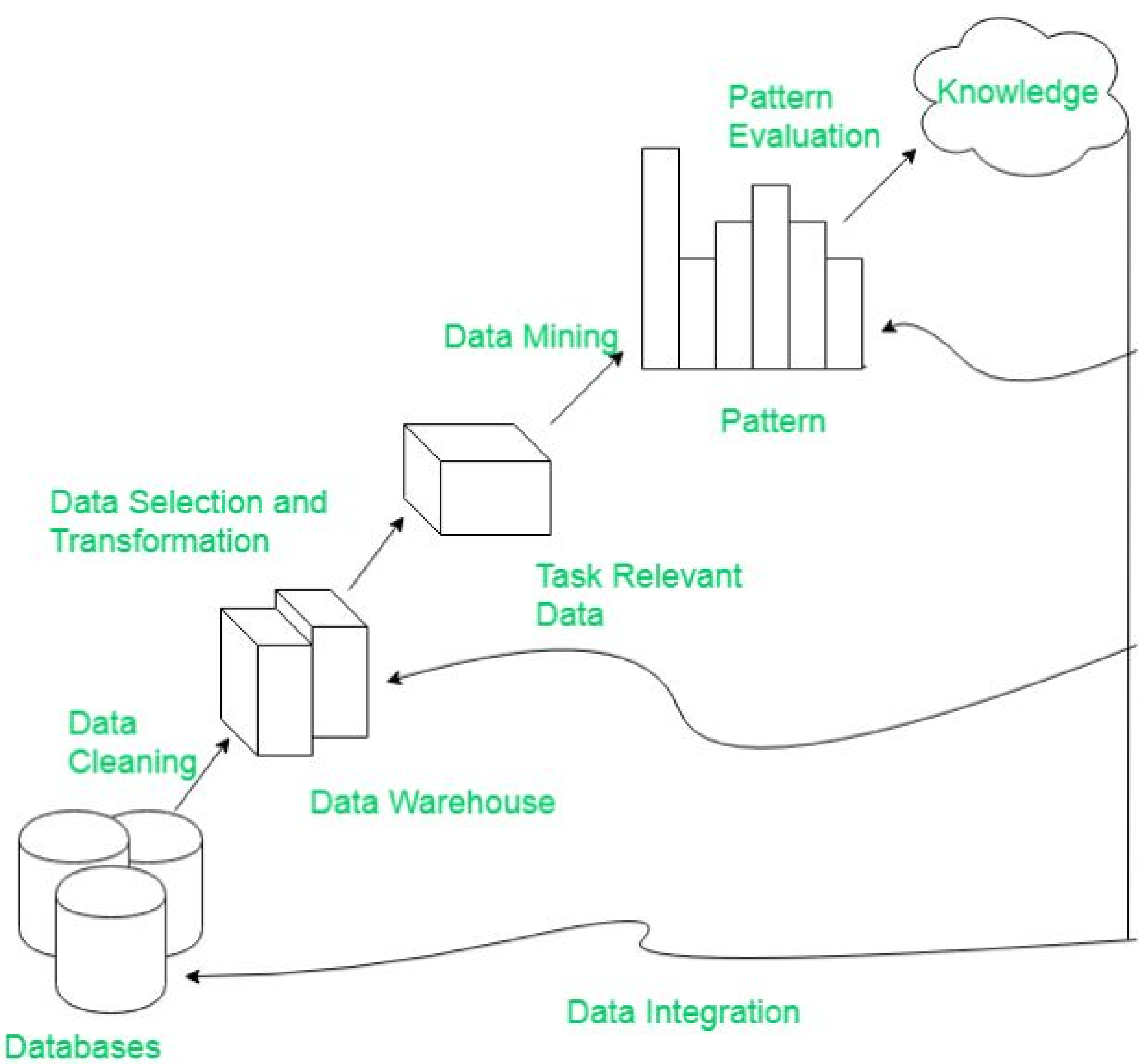
Pattern Evaluation

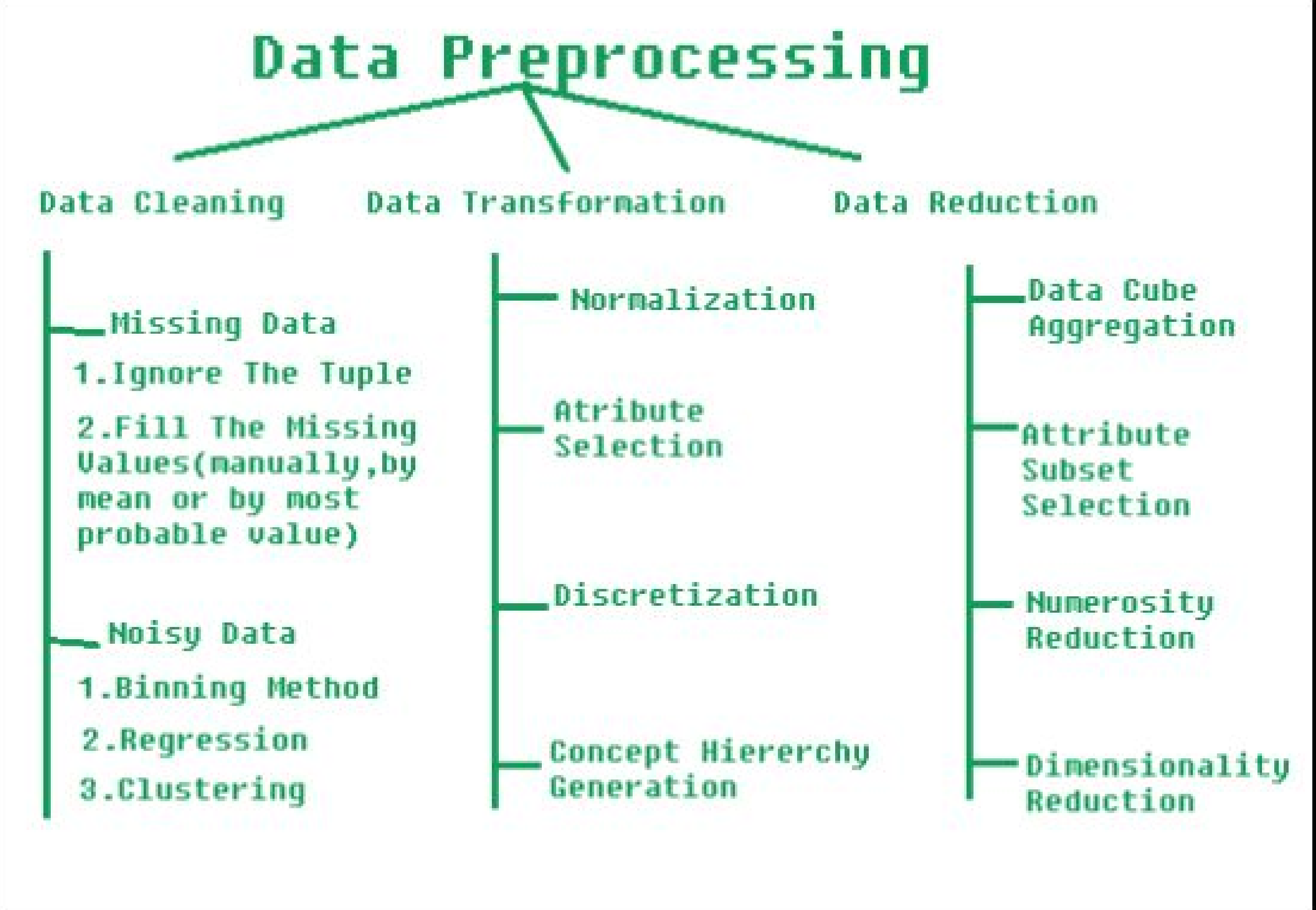
Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures. It find **interestingness score** of each pattern, and uses **summarization** and **Visualization** to make data understandable by user.

Knowledge Representation

This involves presenting the results in a way that is meaningful and can be used to make decisions.




Knowledge
Pattern
Evaluation
Data Mining
Pattern
Data Selection and
Transformation
Task Relevant
Data
Data
Cleaning
Data Warehouse
Data Integration
Databases


Data Preprocessing
Data Cleaning
Data Transformation
Data Reduction
Missing Data
1.Ignore The Tuple
2.Fill The Missing Values(manually,by mean or by most probable value)
Noisy Data
1.Binning Method
2.Regression
3.Clustering
Normalization
Atribute Selection
Discretization
Concept Hiererchy Generation
Data Cube Aggregation
Attribute Subset Selection
Numerosity Reduction
Dimensionality Reduction

Data Preprocessing in Data Mining

Data preprocessing is an important step in the data mining process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

Some common steps in data preprocessing include:

Data preprocessing is an important step in the data mining process that involves cleaning and transforming raw data to make it suitable for analysis. Some common steps in data preprocessing include:

Data Cleaning: This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.

Data Integration: This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling data with different formats, structures, and semantics. Techniques such as record linkage and data fusion can be used for data integration.

Data Transformation: This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, while standardization is used to transform the data to have zero mean and unit variance. Discretization is used to convert continuous data into discrete categories.

Data Reduction: This involves reducing the size of the dataset while preserving the important information. Data reduction can be achieved through techniques such as feature selection and feature extraction. Feature selection involves selecting a subset of relevant features from the dataset, while feature extraction involves transforming the data into a lower-dimensional space while preserving the important information.

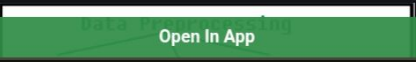
Data Discretization: This involves dividing continuous data into discrete categories or intervals. Discretization is often used in data mining and machine learning algorithms that require categorical data. Discretization can be achieved through techniques such as equal width binning, equal frequency binning, and clustering.

Data Normalization: This involves scaling the data to a common range, such as between 0 and 1 or -1 and 1. Normalization is often used to handle data with different units and scales. Common normalization techniques include min-max normalization, z-score normalization, and decimal scaling.

Data preprocessing plays a crucial role in ensuring the quality of data and the accuracy of the analysis results. The specific steps involved in data preprocessing may vary depending on the nature of the data and the analysis goals.

By performing these steps, the data mining process becomes more efficient and the results become more accurate.

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:

Data reduction is a crucial step in the data mining process that involves reducing the size of the dataset while preserving the important information. This is done to improve the efficiency of data analysis and to avoid overfitting of the model. Some common steps involved in data reduction are:

Feature Selection: This involves selecting a subset of relevant features from the dataset. Feature selection is often performed to remove irrelevant or redundant features from the dataset. It can be done using various techniques such as correlation analysis, mutual information, and principal component analysis (PCA).

Feature Extraction: This involves transforming the data into a lower-dimensional space while preserving the important information. Feature extraction is often used when the original features are high-dimensional and complex. It can be done using techniques such as PCA, linear discriminant analysis (LDA), and non-negative matrix factorization (NMF).

Sampling: This involves selecting a subset of data p ts from the dataset. Sampling is often used to reduce the size of the dataset while preserving t information. It can be done using 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:
Data reduction is a crucial step in the data mining process that involves reducing the size of the dataset while preserving the important information. This is done to improve the efficiency of data analysis and to avoid overfitting of the model. Some common steps involved in data reduction are:

Feature Selection: This involves selecting a subset of relevant features from the dataset. Feature selection is often performed to remove irrelevant or redundant features from the dataset. It can be done using various techniques such as correlation analysis, mutual information, and principal component analysis (PCA).

Feature Extraction: This involves transforming the data into a lower-dimensional space while preserving the important information. Feature extraction is often used when the original features are high-dimensional and complex. It can be done using techniques such as PCA, linear discriminant analysis (LDA), and non-negative matrix factorization (NMF).

Sampling: This involves selecting a subset of data points from the dataset. Sampling is often used to reduce the size of the dataset while preserving the important information. It can be done using techniques such as random sampling, stratified sampling, and systematic sampling.

Clustering: This involves grouping similar data points together into clusters. Clustering is often used to reduce the size of the dataset by replacing similar data points with a representative centroid. It can be done using techniques such as k-means, hierarchical clustering, and density-based clustering.

Compression: This involves compressing the dataset while preserving the important information. Compression is often used to reduce the size of the dataset for storage and transmission purposes. It can be done using techniques such as wavelet compression, JPEG compression, and gzip compression.

Supervised and Unsupervised learning

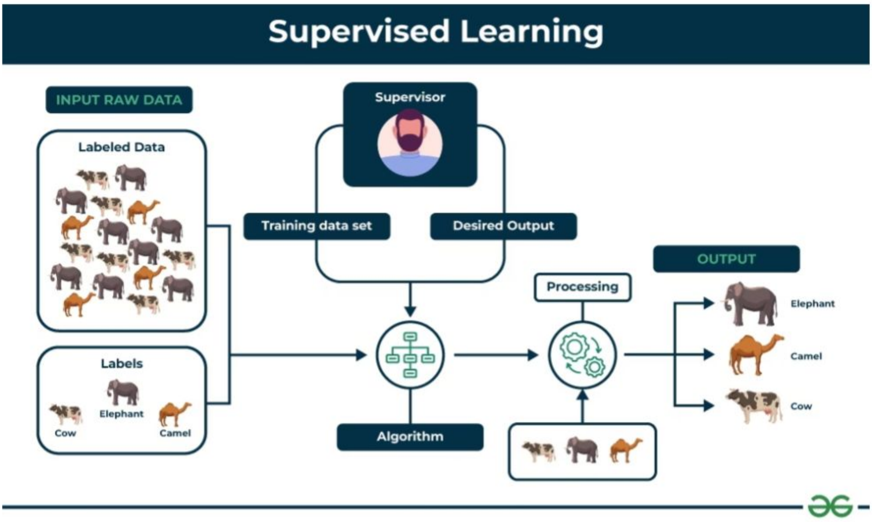
Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Supervised learning and unsupervised learning are two main types of machine learning.

In supervised learning, the machine is trained on a set of labeled data, which means that the input data is paired with the desired output. The machine then learns to predict the output for new input data. Supervised learning is often used for tasks such as classification, regression, and object detection.

In unsupervised learning, the machine is trained on a set of unlabeled data, which means that the input data is not paired with the desired output. The machine then learns to find patterns and relationships in the data. Unsupervised learning is often used for tasks such as clustering, dimensionality reduction, and anomaly detection.

What is Supervised learning?

Supervised learning is a type of machine learning algorithm that learns from labeled data. Labeled data is data that has been tagged with a correct answer or classification.



Key Points:

- Supervised learning involves training a machine from labeled data.
- Labeled data consists of examples with the correct answer or classification.
- The machine learns the relationship between inputs (fruit images) and outputs (fruit labels).
- The trained machine can then make predictions on new, unlabeled data.

Example:

Let's say you have a fruit basket that you want to identify. The m ... first analyze

Types of Supervised Learning

Supervised learning is classified into two categories of algorithms:

- Regression: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.
- Classification: A classification problem is when the output variable is a category, such as “Red” or “blue” , “disease” or “no disease”.

Supervised learning deals with or learns with “labeled” data. This implies that some data is already tagged with the correct answer.

1- Regression

Regression is a type of supervised learning that is used to predict continuous values, such as house prices, stock prices, or customer churn. Regression algorithms learn a function that maps from the input features to the output value.

1- Regression

Regression is a type of supervised learning that is used to predict continuous values, such as house prices, stock prices, or customer churn. Regression algorithms learn a function that maps from the input features to the output value.

Some common regression algorithms include:

- Linear Regression
- Polynomial Regression
- Support Vector Machine Regression
- Decision Tree Regression
- Random Forest Regression

2- Classification

Classification is a type of supervised learning that is used to predict categorical values, such as whether a customer will churn or not, whether an email is spam or not, or whether a medical image shows a tumor or not. Classification algorithms learn a function that

1. Linear Regression
2. Polynomial Regression
3. Stepwise Regression
4. Decision Tree Regression
5. Random Forest Regression
6. Support Vector Regression
7. Ridge Regression
8. Lasso Regression
9. ElasticNet Regression
10. Bayesian Linear Regression

Linear Regression

Linear regression is used for predictive analysis. Linear regression is a linear approach for modeling the relationship between the criterion or the scalar response and the multiple predictors or explanatory variables. Linear regression focuses on the conditional probability distribution of the response given the values of the predictors. For linear regression, there is a danger of overfitting. The formula for linear regression is:

2- Classification

Classification is a type of supervised learning that is used to predict categorical values, such as whether a customer will churn or not, whether an email is spam or not, or whether a medical image shows a tumor or not. Classification algorithms learn a function that maps from the input features to a probability distribution over the output classes.

Some common classification algorithms include:

- Logistic Regression
- Support Vector Machines
- Decision Trees
- Random Forests
- Naive Baye

Evaluating Supervised Learning Models

Evaluating supervised learning models is an important step in ensuring that the model is

Classification Types

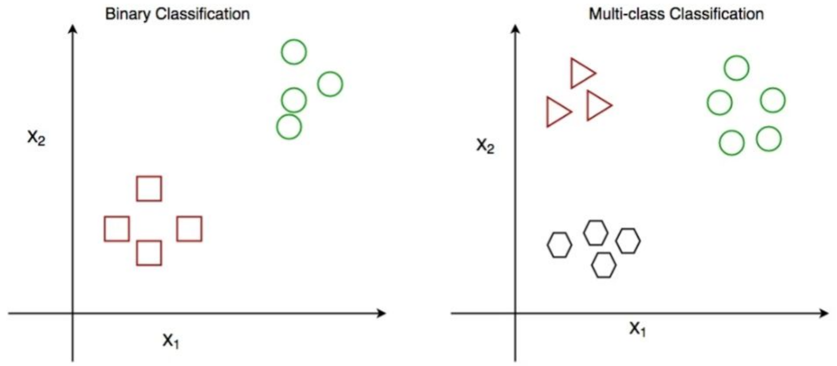
There are two main classification types in machine learning:

Binary Classification

In binary classification, the goal is to classify the input into one of two classes or categories. Example – On the basis of the given health conditions of a person, we have to determine whether the person has a certain disease or not.

Multiclass Classification

In multi-class classification, the goal is to classify the input into one of several classes or categories. For Example – On the basis of data about different species of flowers, we have to determine which specie our observation belongs to.



Binary vs Multi class classification

Other categories of classification involves:

Multi-Label Classification

In, Multi-label Classification the goal is to predict which of several labels a new data point belongs to. This is different from multiclass classification, where each data point can only belong to one class. For example, a multi-label classification algorithm could be used to classify images of animals as belonging to one or more of the categories cat, dog, bird, or fish.

Imbalanced Classification

Imbalanced Classification

In, [Imbalanced Classification]() the goal is to predict whether a new data point belongs to a minority class, even though there are many more examples of the majority class. For example, a medical diagnosis algorithm could be used to predict whether a patient has a rare disease, even though there are many more patients with common diseases.

Classification Algorithms

There are various types of **classifiers algorithms**. Some of them are :

Linear Classifiers

Linear models create a linear decision boundary between classes. They are simple and computationally efficient. Some of the linear **classification** models are as follows:

- Logistic Regression
- Support Vector Machines having kernel =

algorithms. Some of them are :

Linear Classifiers

Linear models create a linear decision boundary between classes. They are simple and computationally efficient. Some of the linear **classification** models are as follows:

- Logistic Regression
- Support Vector Machines having kernel = 'linear'
- Single-layer Perceptron
- Stochastic Gradient Descent (SGD) Classifier

Non-linear Classifiers

Non-linear models create a non-linear decision boundary between classes. They can capture more complex relationships between the input features and the target variable. Some of the non-linear **classification** models are as follows:

- K-Nearest Neighbours

Non-linear Classifiers

Non-linear models create a non-linear decision boundary between classes. They can capture more complex relationships between the input features and the target variable. Some of the non-linear **classification** models are as follows:

- K-Nearest Neighbours
- Kernel SVM
- Naive Bayes
- Decision Tree Classification
- Ensemble learning classifiers:
- Random Forests,
- AdaBoost,
- Bagging Classifier,
- Voting Classifier,
- ExtraTrees Classifier
- Multi-layer Artificial Neural Networks

Learners in Classifications Algorithm

In machine learning, classification learners can also be classified as either "lazy" or

You
application of classification and prediction in supervised learning

ChatGPT
Supervised learning is a type of machine learning where the model is trained on labeled data. The goal is to learn a mapping from inputs to outputs based on the provided examples. Supervised learning is widely used for classification and prediction tasks. Here's an overview of how these two applications work:

1. Classification

Classification is a supervised learning task where the goal is to assign a label from a finite set of categories to an input. Here are some common applications:

Applications:

- **Image Recognition**: Identifying objects within an image (e.g., distinguishing between cats and dogs).
- **Spam Detection**: Classifying emails as spam or not spam.
- **Medical Diagnosis**: Predicting the presence or absence of a disease based on medical data.
- **Sentiment Analysis**: Determining the sentiment of a piece of text (e.g., positive, negative, neutral).
- **Fraud Detection**: Identifying fraudulent transactions.

Example Workflow:

1. **Data Collection**: Gather labeled examples of data. For instance, for image recognition, collect images labeled with the corresponding object.
2. **Preprocessing**: Clean and prepare the data, including normalization, resizing images, and handling missing values.
3. **Model Selection**: Choose a classification algorithm (e.g., Decision Trees, Support Vector Machines, Neural Networks).
4. **Training**: Train the model using the labeled dataset.
5. **Evaluation**: Assess the model's performance using metrics such as accuracy, precision, recall, and F1-score.
6. **Deployment**: Use the trained model to classify new, unseen data.

2. Prediction (Regression)

Prediction in supervised learning often refers to regression tasks where the goal is to predict a continuous value based on input features. Here are some common applications:

Applications:

- **Stock Price Prediction**: Forecasting future stock prices based on historical data.
- **Weather Forecasting**: Predicting future weather conditions using past meteorological data.
- **House Price Prediction**: Estimating the price of a house based on features like location, size, and number of rooms.
- **Demand Forecasting**: Predicting future demand for products or services.
- **Energy Consumption**: Predicting future energy usage based on historical consumption data.

Example Workflow:

1. **Data Collection**: Gather historical data with known continuous outputs (e.g., past stock prices).
2. **Preprocessing**: Clean and prepare the data, including normalization, handling missing values, and feature engineering.
3. **Model Selection**: Choose a regression algorithm (e.g., Linear Regression, Random Forest, Gradient Boosting).
4. **Training**: Train the model using the historical dataset.
5. **Evaluation**: Assess the model's performance using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

1. **Data Collection**: Gather labeled examples of data. For instance, for image recognition, collect images labeled with the corresponding object.
2. **Preprocessing**: Clean and prepare the data, including normalization, resizing images, and handling missing values.
3. **Model Selection**: Choose a classification algorithm (e.g., Decision Trees, Support Vector Machines, Neural Networks).
4. **Training**: Train the model using the labeled dataset.
5. **Evaluation**: Assess the model's performance using metrics such as accuracy, precision, recall, and F1-score.
6. **Deployment**: Use the trained model to classify new, unseen data.

2. Prediction (Regression)

Prediction in supervised learning often refers to regression tasks where the goal is to predict a continuous value based on input features. Here are some common applications:

Applications:

- **Stock Price Prediction**: Forecasting future stock prices based on historical data.
- **Weather Forecasting**: Predicting future weather conditions using past meteorological data.
- **House Price Prediction**: Estimating the price of a house based on features like location, size, and number of rooms.
- **Demand Forecasting**: Predicting future demand for products or services.
- **Energy Consumption**: Predicting future energy usage based on historical consumption data.

Example Workflow:

1. **Data Collection**: Gather historical data with known continuous outputs (e.g., past stock prices).
2. **Preprocessing**: Clean and prepare the data, including normalization, handling missing values, and feature engineering.
3. **Model Selection**: Choose a regression algorithm (e.g., Linear Regression, Random Forest, Gradient Boosting).
4. **Training**: Train the model using the historical dataset.
5. **Evaluation**: Assess the model's performance using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.
6. **Deployment**: Use the trained model to predict future values based on new input data.

Algorithms Used

- **Classification Algorithms**: Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), Neural Networks.
- **Regression Algorithms**: Linear Regression, Polynomial Regression, Ridge Regression, Lasso Regression, Decision Trees, Random Forest, Gradient Boosting Machines (GBM).

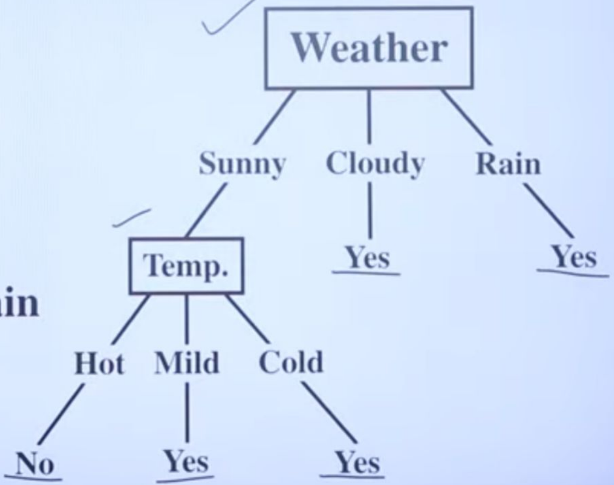
Key Considerations

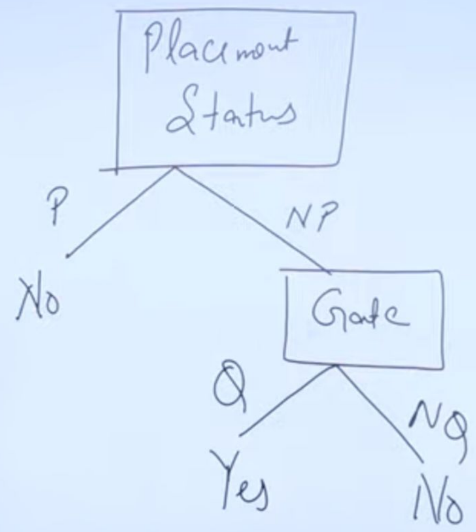
- **Data Quality**: High-quality, labeled data is crucial for both classification and prediction tasks.
- **Feature Selection**: Choosing the right features significantly impacts the performance of the model.
- **Overfitting and Underfitting**: Balance model complexity to avoid overfitting (too complex) and underfitting (too simple).
- **Model Evaluation**: Use appropriate metrics and validation techniques (e.g., cross-validation) to evaluate the model's performance.

By leveraging supervised learning for classification and prediction, organizations can make informed decisions, automate processes, and gain insights from data across various domains.

Decision Tree

- Machine learning algorithm used for both classification and regression tasks.
- Tree Structure
- Decision Nodes
- Leaf Nodes
- Splitting
- Entropy and Information Gain
- [illegible]ning




Placement Status
P
NP
No
Gate
Q
NQ
Yes
No

What is decision tree induction

ChatGPT

Decision tree induction is a popular supervised learning algorithm used for both classification and regression tasks. It involves creating a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Here's a detailed explanation of the process:

Key Concepts of Decision Tree Induction

1. **Decision Tree Structure**:
    - **Root Node**: The topmost node representing the entire dataset, which is split into subsets.
    - **Internal Nodes**: Nodes where the data is split based on feature values.
    - **Leaf Nodes**: Terminal nodes that represent the outcome or class label.
2. **Splitting Criteria**:
    - The process of dividing a node into two or more sub-nodes. The choice of feature and threshold for splitting is critical and depends on the algorithm used (e.g., ID3, C4.5, CART).
3. **Purity Measures**:
    - **Gini Index**: Used in CART (Classification and Regression Trees) to measure the impurity of a node.
    - **Information Gain**: Used in ID3 and C4.5, it measures the reduction in entropy before and after a split.
    - **Chi-square**: Used in CHAID (Chi-squared Automatic Interaction Detector) for statistical tests of independence.

Steps in Decision Tree Induction

1. **Start with the Root Node**: The root node represents the entire dataset. The algorithm starts here and evaluates each feature to determine the best split.
2. **Select the Best Feature to Split**:
    - Evaluate each feature and corresponding split point to determine the best feature to split the data.
    - Use a splitting criterion like Gini Index, Information Gain, or others to decide the best feature and split point.
3. **Split the Node**:
    - Divide the dataset into subsets based on the selected feature and split point.
    - Create child nodes for the resulting subsets.
4. **Repeat the Process**:
    - Recursively apply the same splitting process to each child node.
    - Continue splitting nodes until a stopping criterion is met (e.g., maximum depth, minimum samples per leaf, or no further improvement in purity).
5. **Assign Class Labels**:
    - Once the splitting process stops, assign class labels to the leaf nodes based on the majority class or average value (for regression).

Advantages and Disadvantages

Advantages:

Day	Weather	Temperature	Humidity	Wind	Play Football?
Day 1	Sunny	Hot	High	Weak	No
Day 2	Sunny	Hot	High	Strong	No
Day 3	Cloudy	Hot	High	Weak	Yes
Day 4	Rain	Mild	High	Weak	Yes
Day 5	Rain	Cool	Normal	Weak	Yes
Day 6	Rain	Cool	Normal	Strong	No
Day 7	Cloudy	Cool	Normal	Strong	Yes
Day 8	Sunny	Mild	High	Weak	No
Day 9	Sunny	Cool	Normal	Weak	Yes
Day 10	Rain	Mild	Normal	Weak	Yes
Day [illegible]	Sunny	Mild	Normal	Strong	Yes
[illegible] 12	Cloudy	Mild	High	Strong	Yes
Day 13	Cloudy	Hot	Normal	Weak	Yes
Day 14	[illegible]	Mild	High	Strong	No

SUBSCRIBE

Calculate IG of Weather

Step1: Entropy of entire dataset

$$S\{+9,-5\} = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

Step2: Entropy of all attributes:

- Entropy of Sunny $\{+2,-3\} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$
- Entropy of Cloudy$\{+4,-0\} = -\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4} = 0$
- Entropy of Rain$\{+3,-2\} = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.97$

Inform[illegible]Gain= Entropy(whole data) $-\frac{5}{14}\text{Ent(S)} - \frac{4}{14}\text{Ent(C)} - \frac{5}{14}\text{Ent(R)}$

$$= 0.246$$

Calculate IG of Temperature

- **Step1: Entropy of entire dataset**

$$S\{+9,-5\} = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = \underbrace{0.94}$$

- **Step2: Entropy of all attributes:**
- **Entropy of Hot** $\{+2,-2\} = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1.0$
- **Entropy of Mild**$\{+4,-2\} = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.91$
- **Entropy of Cold**$\{+3,-1\} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{3}\log_2\frac{1}{3} = 0.81$
- **Information Gain= Entropy(whole data)** $- \frac{4}{14}\text{Ent(H)} - \frac{6}{14}\text{Ent(M)} - \frac{4}{14}\text{Ent(C)}$

$$= 0.029$$

Calculate IG of Humidity

- **Step1: Entropy of entire dataset**

$$S\{+9,-5\} = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

- **Step2: Entropy of all attributes:**
- **Entropy of High** $\{+3,-4\} = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.98$
- **Entropy of Normal** $\{+6,-1\} = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0.59$

- **Information Gain= Entropy(whole data)** $-\frac{7}{14}\text{Ent}(H) - \frac{7}{14}\text{Ent}(N)$

$$= 0.15$$

Calculate IG of Wind

- Step1: Entropy of entire dataset

$$S\{+9,-5\} = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$

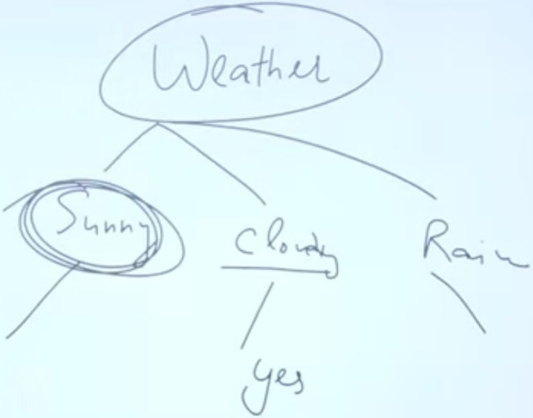
- Step2: Entropy of all attributes:
- Entropy of Strong $\{+3,-3\} = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1.0$
- Entropy of Normal $\{+6,-2\} = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} = 0.81$

- Information Gain= Entropy(whole data) $-\frac{6}{14}$Ent(S) $-\frac{8}{14}$Ent(W)

$= 0.0478$

- Gain (S, Weather) = 0.246
- Gain (S, Temp) = 0.029
- Gain (S, Humidity) = 0.15
- Gain (S, Wind) = 0.0478

- (S, Weather) = 0.246
- S, Temp) = 0.029
- S, Humidity) = 0.15
- S, Wind) = 0.0478



Association Rule Mining: (Rules)

"ARM", Also Called as Market Basket Analysis (MBA) and Affinity Analysis.

↳ Set of items in a transaction is Called Market Basket.

↳ Mostly used in RETAIL.

↳ if 'A' then 'B' $\{A \Rightarrow B\}$
- 'A', 'B' → Product
- A → antecedent
- B → Consequent.

Support: (S). Percentage (%) of transactions (T) that contains both 'A' and 'B'.

$(\text{"A"} \Rightarrow \text{"B"}) = P(A \cap B)$ } measures frequency of association.

Confidence: (C). In a transaction set 'T' if 'C' is the % of times 'B' is present in all the transactions containing 'A'.

Conditional Probability $\left\{ C = P(B/A) = \dfrac{P(A \cap B)}{P(A)} \right\}$ ~~Strength~~ (Strength) of association

Parameters:-

(i) Finding all items that appears frequently in transaction. } min. Support Count.

(ii) Finding Strong associations among frequent items } Confidence.

Data Warehouse and Data Mining – Video Lecture Series (For B.Tech, MCA, M.Tech)

Problems in ARM:-

i) Levels of frequency of appearance determination.

ii) Finding strong associations among frequent items.

Functions of ARM:-

i) Finding set of items that has significant impact on business.

ii) Collating info^n from numerous tr^x.

iii) Generating rules from counts in tr^x.

Strengths of ARM:-

i) Easy interpretation.

ii) Easy to start

iii) Flexible data formats

iv) Simplicity.

(1,2,3,4)

(1,2), (1,3), (1,2,3) ...

Weakness:-

i) Exponential Growth in Computations

ii) Lumping

iii) Rule Selection

iv) Rare Items } frequent items

Data Warehouse and Data Mining – Video Lecture Series (For B.Tech, MCA, M.Tech)

Apriori Algorithm: Idea is to generate Candidate itemsets of a given Size and then Scan dataset to check if their Counts are really large. The Process is iterative.

(i) All Singleton itemsets are Candidates in the first pass. Any items with less than specified Support Value is eliminated.

(ii) Two member Candidate itemsets. {Numericals}

(iii) Three " " " " ".

(iv) Frequent itemsets Constitutes Set of frequent itemsets.

(v) Generate Association Rules which have Confidence Values greater than or equal to Specified min. Confidence.

Eg:-

Tid	Items
1	2,3
2	1,3,5
3	1,2,4
4	2,3

min Support = 2

Items	Support
1	→ 2
2	→ 3
3	→ 3
4	→ 1
5	→ 1

eliminated.

Itemsets	Support
{1,2}	→ 1
{1,3}	→ 1
{2,3}	→ 2

eliminated.

{2,3} → 2

Association Rule Learning

Association rule learning is a type of unsupervised learning technique that checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable. It tries to find some interesting relations or associations among the variables of dataset. It is based on different rules to discover the interesting relations between variables in the database.

The association rule learning is one of the very important concepts of machine learning, and it is employed in **Market Basket analysis, Web usage mining, continuous production, etc.** Here market basket analysis is a technique used by the various big retailer to discover the associations between items. We can understand it by taking an example of a supermarket, as in a supermarket, all products that are purchased together are put together.

For example, if a customer buys bread, he most likely can also buy butter, eggs, or milk, so these products are stored within a shelf or mostly nearby. Consider the below diagram:

- Support
- Confidence
- Lift

Let's understand each of them:

Support

Support is the frequency of A or how frequently an item appears in the dataset. It is defined as the fraction of the transaction T that contains the itemset X. If there are X datasets, then for transactions T, it can be written as:

Supp(X)= $\frac{Freq(X)}{T}$

Confidence

Confidence indicates how often the rule has been found to be true. Or how often the items X and Y occur together in the dataset when the occurrence of X is already given. It is the ratio of the transaction that contains X and Y to the number of records that contain X.

Confidence= $\frac{Freq(X,Y)}{Freq(X)}$

Lift

It is the strength of any rule, which can be

Confidence= $\frac{Freq(X,Y)}{Freq(X)}$

Lift

It is the strength of any rule, which can be defined as below formula:

Lift= $\frac{Supp(X,Y)}{Supp(X)\times Supp(Y)}$

It is the ratio of the observed support measure and expected support if X and Y are independent of each other. It has three possible values:

- If **Lift= 1**: The probability of occurrence of antecedent and consequent is independent of each other.
- **Lift>1**: It determines the degree to which the two itemsets are dependent to each other.
- **Lift<1**: It tells us that one item is a substitute for other items, which means one item has a negative effect on another.

Types of Association Rule Lerning

Association rule learning can be divided into three algorithms:

Apriori Algorithm

This algorithm uses frequent datasets to generate association rules. It is designed to negative effect on another.

Types of Association Rule Lerning

Association rule learning can be divided into three algorithms:

Apriori Algorithm

This algorithm uses frequent datasets to generate association rules. It is designed to work on the databases that contain transactions. This algorithm uses a breadth-first search and Hash Tree to calculate the itemset efficiently.

It is mainly used for market basket analysis and helps to understand the products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.

Eclat Algorithm

Eclat algorithm stands for **Equivalence Class Transformation**. This algorithm uses a depth-first search technique to find frequent itemsets in a transaction database. It performs faster execution than Apriori Algorithm.

F-P Growth Algorithm

The F-P growth algorithm stands for **Frequent Pattern**, and it is the improved version of the Apriori Algorithm. It represents the database in itemset efficiently.

It is mainly used for market basket analysis and helps to understand the products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.

Eclat Algorithm

Eclat algorithm stands for **Equivalence Class Transformation**. This algorithm uses a depth-first search technique to find frequent itemsets in a transaction database. It performs faster execution than Apriori Algorithm.

F-P Growth Algorithm

The F-P growth algorithm stands for **Frequent Pattern**, and it is the improved version of the Apriori Algorithm. It represents the database in the form of a tree structure that is known as a frequent pattern or tree. The purpose of this frequent tree is to extract the most frequent patterns.

Applications of Association Rule Learning

It has various applications in machine learning and data mining. Below are some popular applications of association rule learning:

- **Market Basket Analysis:** It is one of the popular examples and applications of association rule mining. This technique is commonly used by big retailers to determine the association between items.
- **Medical Diagnosis:** With the help of association rules, patients can be cured easily, as it helps in identifying the probability of illness for a particular disease.
- **Protein Sequence:** The association rules help in determining the synthesis of artificial Proteins.
- It is also used for the **Catalog Design** and **Loss-leader Analysis** and many more other applications.

Association Rules Exercise

- Here are a dozen sales transactions.
- The objective is to use this transaction data to find affinities between products, that is, which products sell together often.
- The support level will be set at 33 percent; the confidence level will be set at 50 percent.

Association Rules Exercise

Transactions List

1	Milk	Egg	Bread	Butter
2	Milk	Butter	Egg	Ketchup
3	Bread	Butter	Ketchup	
4	Milk	Bread	Butter	
5	Bread	Butter	Cookies	
6	Milk	Bread	Butter	Cookies
7	Milk	Cookies		
8	Milk	Bread	Butter	
9	Bread	Butter	Egg	Cookies
10	Milk	Butter	Bread	
11	Milk	Bread	Butter	
12	Milk	Bread	Cookies	Ketchup

Association Rules Exercise

- Here are a dozen sales transactions.
- The objective is to use this transaction data to find affinities between products, that is, which products sell together often.
- The support level will be set at 33 percent; the confidence level will be set at 50 percent.

Association Rules Exercise

$$Rule: \quad X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

Transactions List

1	Milk	Egg	Bread	Butter
2	Milk	Butter	Egg	Ketchup
3	Bread	Butter	Ketchup	
4	Milk	Bread	Butter	
5	Bread	Butter	Cookies	
6	Milk	Bread	Butter	Cookies
7	Milk	Cookies		
8	Milk	Bread	Butter	
9	Bread	Butter	Egg	Cookies
10	Milk	Butter	Bread	
11	Milk	Bread	Butter	
12	Milk	Bread	Cookies	Ketchup

1-item Sets	Frequency
Milk	9
Bread	10
Butter	10
Egg	3
Ketchup	3
Cookies	5

Frequent 1-item Sets	Frequency
Milk	9
Bread	10
Butter	10
Cookies	5

Transactions List

1	Milk	Egg	Bread	Butter
2	Milk	Butter	Egg	Ketchup
3	Bread	Butter	Ketchup	
4	Milk	Bread	Butter	
5	Bread	Butter	Cookies	
6	Milk	Bread	Butter	Cookies
7	Milk	Cookies		
8	Milk	Bread	Butter	
9	Bread	Butter	Egg	Cookies
10	Milk	Butter	Bread	
11	Milk	Bread	Butter	
12	Milk	Bread	Cookies	Ketchup

2-item Sets	Frequency
Milk, Bread	7
Milk, Butter	7
Milk, Cookies	3
Bread, Butter	9
Butter, Cookies	3
Bread, Cookies	4

Frequent 2-item Sets	Frequency
Milk, Bread	7
Milk, Butter	7
Bread, Butter	9
Bread, Cookies	4

Transactions List

1	Milk	Egg	Bread	Butter
2	Milk	Butter	Egg	Ketchup
3	Bread	Butter	Ketchup	
4	Milk	Bread	Butter	
5	Bread	Butter	Cookies	
6	Milk	Bread	Butter	Cookies
7	Milk	Cookies		
8	Milk	Bread	Butter	
9	Bread	Butter	Egg	Cookies
10	Milk	Butter	Bread	
11	Milk	Bread	Butter	
12	Milk	Bread	Cookies	Ketchup

Milk, Bread, Butter, Cookies

3-item Sets	Frequency
Milk, Bread, Butter	6
Milk, Bread, Cookies	1
Bread, Butter, Cookies	3
Milk, Butter, Cookies	2

Frequent 3-item Sets	Frequency
Milk, Bread, Butter	6

Association Rule Mining - Subset Creation

- Frequent 3-Item Set = I => {Milk, Bread, Butter}

- Non-Empty subset are
  - {{Milk}, {Bread}, {Butter}, {Milk, Bread}, {Milk, Butter}, {Bread, Butter}}

- How to form Association Rule…?
  - For every non-empty subset S of I, the association rule is,
    - **S → (I-S)**
    - **If support(I) / support(S) >= min_confidence**

Association Rule Mining - Subset Creation

- Non-Empty subset are
  - {{Milk}, {Bread}, {Butter}, {Milk, Bread}, {Milk, Butter}, {Bread, Butter}}
  - Min_Support = 30% and Min_Confidence = 60%

- Rule 1: {Milk} → {Bread, Butter} {S=50%, C=66.67%}
  - Support = 6/12 = 50%
  - Confidence = Support (Milk, Bread, Butter)/Support(Milk) = $\frac{6/12}{9/12}$ = 6/9 = 66.67% > 60%
  - Valid

- Rule 2: {Bread} → {Milk, Butter} {S=50%, C=60%}
  - Support = 6/12 = 50%
  - Confidence = Support (Milk, Bread, Butter)/Support(Bread) = 6/10 = 60% >= 60%
  - Valid

Association Rule Mining - Subset Creation

- Non-Empty subset are
  - {{Milk}, {Bread}, {Butter}, {Milk, Bread}, {Milk, Butter}, {Bread, Butter}}
  - Min_Support = 30% and Min_Confidence = 60%

- Rule 3: {Butter} → {Milk, Bread} {S=50%, C=60%}
  - Support = 6/12 = 50%
  - Confidence = Support (Milk, Bread, Butter)/Support(Butter) = 6/10 = 60%>=60
  - Valid

- Rule 4: {Milk, Bread} → {Butter} {S=50%, C=85.7%}
  - Support = 6/12 = 50%
  - Confidence = Support (Milk, Bread, Butter)/Support(Milk, Bread) = 6/7 = 85.7% > 60%
  - Valid

Data Warehouse and [Mumbai Univ, Pune Univ, GTU, PTU,
Data Mining Lecture Series [UPTU, GGSIPU and other Univ.]

Easy Engineering Classes – Free YouTube Coaching

For Engineering Students of GGSIPU, UPTU and Other Universities, Colleges of India

Solved Question on FP-GROWTH Algorithm

Ques.) Generate FP-Tree for the following Transaction Data Set. [Minimum Support = 30%]

TR. Id.	Items
1	E, A, D, B
2	D, A, C, E, B
3	C, A, B, E
4	B, A, D
5	D
6	D, B
7	A, D, E
8	B, C

min. no. of Tr^n = 2.4 ⇒ 3

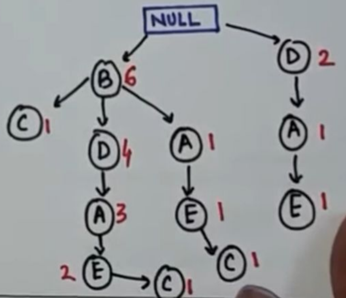
items	Frequency	Priority
A →	5 →	3
B →	6 →	1
C →	3 →	5
D →	6 →	2
E →	4 →	4

Lower Priority no. means High Priority.

Order the items According to the priority.

Tr. Id	Items	Ordered Items
1	E, A, D, B	B, D, A, E ✓
2	D, A, C, E, B	B, D, A, E, C ✓
3	C, A, B, E	B, A, E, C ✓
4	B, A, D	B, D, A ✓
5	D	D ✓
6	D, B	B, D ✓
7	A, D, E	D, A, E ✓
8	B, C	B, C

B:1, 2, 3, 4, 5, 6
D:1, 2, 3, 4
A:1, 2, 3
E:1, 2
C:1
A:1 → 3
E:1
C:1
D:1, 2
A:1
E:1
C:1



Data Mining – Cluster Analysis

Last Updated : 01 Feb, 2023

INTRODUCTION:

Cluster analysis, also known as clustering, is a method of data mining that groups similar data points together. The goal of cluster analysis is to divide a dataset into groups (or clusters) such that the data points within each group are more similar to each other than to data points in other groups. This process is often used for exploratory data analysis and can help identify patterns or relationships within the data that may not be immediately obvious. There are many different algorithms used for cluster analysis, such as k-means, hierarchical clustering, and density-based clustering. The choice of algorithm will depend on the specific requirements of the analysis and the nature of the data being analyzed.

Cluster Analysis is the process to find similar groups of objects in order to form clusters. It is an unsupervised machine learning-based algorithm that acts on unlabelled data. A group of data points would comprise together to form a cluster in which all the objects would belong to the same group.

The given data is divided into different groups by combining similar objects into a group. This group is nothing but a cluster. A cluster is nothing but a collection of similar data which is grouped together.

For example, consider a dataset of vehicles given in which it contains information about different vehicles like cars, buses, bicycles, etc. As it is unsupervised learning there are no class labels like Cars, Bikes, etc for all the vehicles, all the data is combined and is not in a structured manner.

Properties of Clustering :

1. **Clustering Scalability:** Nowadays there is a vast amount of data and should be dealing with huge databases. In order to handle extensive databases, the clustering algorithm should be scalable. Data should be scalable, if it is not scalable, then we can't get the appropriate result which would lead to wrong results.

2. **High Dimensionality:** The algorithm should be able to handle high dimensional space along with the data of small size.

3. **Algorithm Usability with multiple data kinds:** Different kinds of data can be used with algorithms of clustering. It should be capable of dealing with different types of data like discrete, categorical and interval-based data, binary data etc.

4. **Dealing with unstructured data:** There would [illegible] tabases that

There would be some databases that contain missing values, and noisy or erroneous data. If the algorithms are sensitive to such data then it may lead to poor quality clusters. So it should be able to handle unstructured data and give some structure to the data by organising it into groups of similar data objects. This makes the job of the data expert easier in order to process the data and discover new patterns.

5. **Interpretability:** The clustering outcomes should be interpretable, comprehensible, and usable. The interpretability reflects how easily the data is understood.

Clustering Methods:

The clustering methods can be classified into the following categories:

- Partitioning Method

Clustering Methods:

The 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: It is used to make partitions on the data in order to form clusters. If "n" partitions are done on "p" objects of the database then each partition is represented by a cluster and $n < p$. The two conditions which need to be satisfied with this Partitioning Clustering Method are:

- One objective should only belong to only one group.
- There should be no group without even a single purpose.

Applications Of Cluster Analysis:

- It is widely used in image processing, data analysis, and pattern recognition.
- It helps marketers to find the distinct groups in their customer base and they can characterize their customer groups by using purchasing patterns.
- It can be used in the field of biology, by deriving animal and plant taxonomies and identifying genes with the same capabilities.
- It also helps in information discovery by classifying documents on the web.

Advantages of Cluster Analysis:

1. It can help identify patterns and relationships within a dataset that may not be immediately obvious.

Advantages of Cluster Analysis:

1. It can help identify patterns and relationships within a dataset that may not be immediately obvious.

2. It can be used for exploratory data analysis and can help with feature selection.

3. It can be used to reduce the dimensionality of the data.

4. It can be used for anomaly detection and outlier identification.

5. It can be used for market segmentation and customer profiling.

Disadvantages of Cluster Analysis:

1. It can be sensitive to the choice of initial con[illegible] the number of

Disadvantages of Cluster Analysis:

1. It can be sensitive to the choice of initial conditions and the number of clusters.

2. It can be sensitive to the presence of noise or outliers in the data.

3. It can be difficult to interpret the results of the analysis if the clusters are not well-defined.

4. It can be computationally expensive for large datasets.

5. The results of the analysis can be affected by the choice of clustering algorithm used.

6. It is important to note that the success of cluster analysis depends on the data, the goals of the analysis, and the ability of the analyst to interpret the results.

Partitioning Method (K-Mean) in Data Mining

Last Updated : 08 Dec, 2022

Partitioning Method: This clustering method classifies the information into multiple groups based on the characteristics and similarity of the data. Its the data analysts to specify the number of clusters that has to be generated for the clustering methods. In the partitioning method when database(D) that contains multiple(N) objects then the partitioning method constructs user-specified(K) partitions of the data in which each partition represents a cluster and a particular region. There are many algorithms that come under partitioning method some of the popular ones are K-Mean, PAM(K-Medoids), CLARA algorithm (Clustering Large Applications) etc. In this article, we will be seeing the working of K Mean algorithm in detail. **K-Mean (A centroid based Technique):** The K means algorithm takes the input parameter K from the user and partitions the dataset containing N objects into K clusters so that resulting similarity among the data objects inside the group (intracluster) is high but the similarity of data objects with the data objects from outside the cluster is low (intercluster). The similarity of the cluster is determined with respect to the mean value of the cluster. It is a type of square error algorithm. At the start randomly k objects from the dataset are chosen in which each of the objects represents a cluster mean(centre). For the rest of the data objects, they are assigned to the nearest cluster based on their distance from the cluster mean. The new mean of each of the cluster is then calculated with the added data objects. **Algorithm: K mean:**

```
Input:
K: The number of clusters in which the dataset has to be divided
D: A dataset containing N number of objects

Output:
A dataset of K clusters
```

Method:

1. Randomly assign K objects from the dataset(D) as cluster centres(C)
2. (Re) Assign each object to which object is most similar based upon mean values.
3. Update Cluster means, i.e., Recalculate the mean of each cluster with the updated values.
4. Repeat Step 2 until no change occurs.

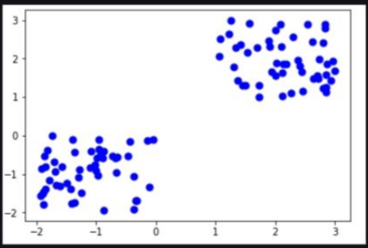


Figure – K-mean Clustering**Flowchart:**

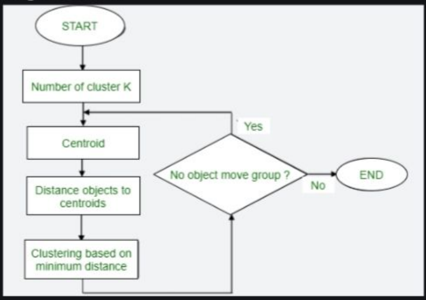


Figure – K-mean Clustering**Example:** Suppose we want to group the visitors to a website using just their age as follows:

Hierarchical Clustering in Data Mining

Last Updated : 12 Dec, 2023

A **Hierarchical clustering** method works via grouping data into a tree of clusters. Hierarchical clustering begins by treating every data point as a separate cluster. Then, it repeatedly executes the subsequent steps:

1. Identify the 2 clusters which can be closest together, and
2. Merge the 2 maximum comparable clusters. We need to continue these steps until all the clusters are merged together.

In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called Dendrogram (A Dendrogram is a tree-like diagram that statistics the sequences of merges or splits) graphically represents this hierarchy and is an inverted tree that describes the order in which factors are merged (bottom-up view) or clusters are broken up (top-down view).

What is Hierarchical Clustering?

Hierarchical clustering is a method of cluster analysis in data mining that creates a hierarchical representation of the clusters in a dataset. The method starts by treating each data point as a separate cluster and then iteratively combines the closest clusters until a stopping criterion is reached. The result of hierarchical clustering is a tree-like structure, called a dendrogram, which illustrates the hierarchical relationships among the clusters.

Hierarchical clustering has several advantages over other clustering methods

- The ability to handle non-convex clusters and clusters of different sizes and densities.
- The ability to handle missing data and noisy data.
- The ability to reveal the hierarchical structure of the data, which can be useful for understanding the relationships among the clusters.

Drawbacks of Hierarchical Clustering

- The need for a criterion to stop the clustering process and determine the final number of clusters.
- The computational cost and memory requirements of the method can be high, especially for large datasets.
- The results can be sensitive to the initial conditions, linkage criterion, and distance metric used. In summary, Hierarchical clustering is a method of data mining that groups similar data points into clusters by creating a hierarchical structure of the clusters.
- This method can handle different types of data and reveal the relationships among the clusters. However, it can have high computational cost and results can be sensitive to some conditions.

Types of Hierarchical Clustering

Basically, there are two types of hierarchical Clustering:

1. Agglomerative Clustering
2. Divisive clustering

DBSCAN Full Form

Last Updated : 16 Dec, 2021

DBSCAN stands for **Density-Based Spatial Clustering of Applications with Noise**.
It is a popular unsupervised learning method used for model construction and machine learning algorithms. It is a clustering method utilized for separating high-density clusters from low-density clusters. It divides the data points into many groups so that points lying in the same group will have the same properties. It was proposed by Martin Ester, Hans-Peter Kriegel, Jorg Sander, and Xiaowei Xu in 1996.
DBSCAN is designed for use with databases that can accelerate region queries. It can not cluster data sets with large differences in their densities.

Characteristics

- It identifies clusters of any shape in a data set, it means it can detect arbitrarily shaped clusters.
- It is based on intuitive notions of clusters and noise.
- It is very robust in detection of outliers in data set
- It requires only two points which are very insensitive to the order of occurrence of the points in data set

Advantages

- Specification of number of clusters of data in the data set is not required.
- It can find any shape cluster even if the cluster is surrounded by any other cluster.
- It can easily find outliers in data set.
- It is not much sensitive to noise, it means it is noise tolerant.
- It is the second most used clustering method after K-means.

Disadvantages

- The quality of the result depends on the distance measure used in the regionQuery function.
- Border points may go in any cluster depending on the processing order so it is not completely deterministic.
- It can be expensive when cost of computation of nearest neighbor is high.
- It can be slow in execution for higher dimension.
- Adaptability of variation in local density is less.

Next Article
DBS Full Form

Cluster Evaluation

- "Clusters can be evaluated with "internal" as well as "external" measures
  - Internal measures are related to the inter/intra cluster distance
    - A good clustering is one where
      - » (Intra-cluster distance) the sum of distances between objects in the same cluster are minimized,
      - » (Inter-cluster distance) while the distances between different clusters are maximized
      - » Objective to minimize: F(Intra,Inter)
  - External measures are related to how representative are the current clusters to "true" classes. Measured in terms of purity, entropy or F-measure
    - Note that in real world, you often *don't know* what the true classes are. (This is why clustering is called unsupervised learning)

Clustering: Application Examples

- Biology: taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- Information retrieval: document clustering
- Land use: Identification of areas of similar land use in an earth observation database
- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults
- Climate: understanding earth climate, find patterns of atmospheric and ocean
- Economic Science: market resarch

Types of Outliers in Data Mining

Last Updated : 17 Apr, 2024

Outlier is a data object that deviates significantly from the rest of the data objects and behaves in a different manner. They can be caused by measurement or execution errors. The analysis of outlier data is referred to as outlier analysis or outlier mining.

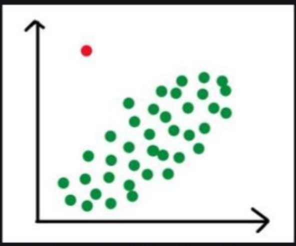
An outlier cannot be termed as a noise or error. Instead, they are suspected of not being generated by the same method as the rest of the data objects.

Outliers are of three types, namely –

1. Global (or Point) Outliers
2. Collective Outliers
3. Contextual (or Conditional) Outliers

1. Global Outliers

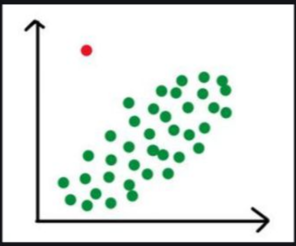
1. Definition: Global outliers are data points that deviate significantly from the overall distribution of a dataset.
2. Causes: Errors in data collection, measurement errors, or truly unusual events can result in global outliers.
3. Impact: Global outliers can distort data analysis results and affect machine learning model performance.
4. Detection: Techniques include statistical methods (e.g., z-score, Mahalanobis distance), machine learning algorithms (e.g., isolation forest, one-class SVM), and data visualization techniques.
5. Handling: Options may include removing or correcting outliers, transforming data, or using robust methods.
6. Considerations: Carefully considering the impact of global outliers is crucial for accurate data analysis and machine learning model outcomes.



The red data point is a global outlier.

2. Collective Outliers

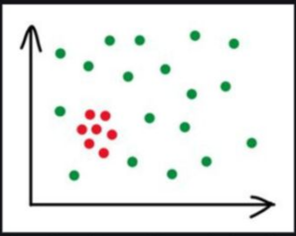
1. Definition: Collective outliers are groups of data points that collectively deviate significantly from the overall distribution of a dataset.
2. Characteristics: Collective outliers may not be outliers when considered individually, but as a group, they exhibit unusual behavior.
3. Detection: Techniques for detecting collective outliers include clustering algorithms, density-based methods, and subspace-based appro...



The red data point is a global outlier.

2. Collective Outliers

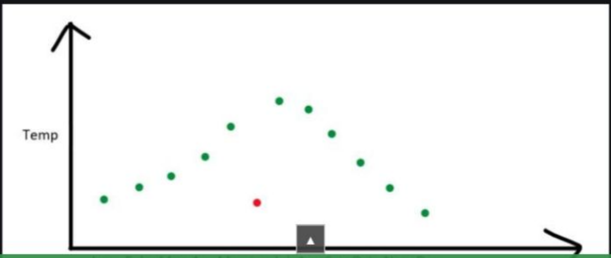
1. Definition: Collective outliers are groups of data points that collectively deviate significantly from the overall distribution of a dataset.
2. Characteristics: Collective outliers may not be outliers when considered individually, but as a group, they exhibit unusual behavior.
3. Detection: Techniques for detecting collective outliers include clustering algorithms, density-based methods, and subspace-based approaches.
4 Impact: Collective outliers can represent interesting patterns or anomalies in data that may require special attention or further investigation.
5. Handling: Handling collective outliers depends on the specific use case and may involve further analysis of the group behavior, identification of contributing factors, or considering contextual information.
6. Considerations: Detecting and interpreting collective outliers can be more complex than individual outliers, as the focus is on group behavior rather than individual data points. Proper understanding of the data context and domain knowledge is crucial for effective handling of collective outliers.



The red data points as a whole are collective outliers.

3. Contextual Outliers

1. Definition: Contextual outliers are data points that deviate significantly from the expected behavior within a specific context or subgroup.
2. Characteristics: Contextual outliers may not be outliers when considered in the entire dataset, but they exhibit unusual behavior within a specific context or subgroup.
3. Detection: Techniques for detecting contextual outliers include contextual clustering, contextual anomaly detection, and context-aware machine learning approaches.
4. Contextual Information: Contextual information such as time, location, or other relevant factors are crucial in identifying contextual outliers.
5. Impact: Contextual outliers can represent unusual or anomalous behavior within a specific context, which may require further investigation or attention.
6. Handling: Handling contextual outliers may involve considering the contextual information, contextual normalization or transformation of data, or using context-specific models or algorithms.
7. Considerations: Proper understanding of the context and domain-specific knowledge is crucial for accurate detection and interpretation of contextual outliers, as they may vary based on the specific context or subgroup being considered.



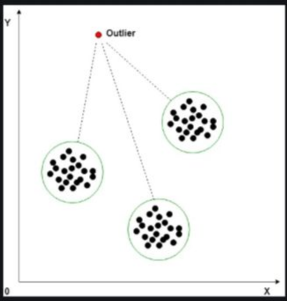
Challenges of Outlier Detection in Data Mining

Last Updated : 01 Mar, 2024

Outlier Detection means finding out the data objects whose properties and behaviour are different from the rest of the objects in the cluster or the data sets. Outlier Detection is the process of finding the outliers from the normal objects. It is essential to perform the Outlier Detection during the data preprocessing. Outliers highly affect the performance of the classification and clustering models. There are many outlier detection methods in data mining. Some of them are as follows:

- Proximity-based methods
- Grid-based methods
- Distance-based methods
- Clustering-based methods

There are a few challenges while applying these outlier detection methods.



For more details please refer to the Types of Outliers article.

The challenges of outlier detection methods in data mining are listed below.

- **Modeling normal outliers effectively:** The quality of Outlier detection depends on the modeling of normal (which are not an outlier) objects. Often, building a model for finding the data normality is very challenging and maybe impossible because it is hard to determine all the behavioral properties of the normal objects. It is difficult to predict the border between normal outliers and abnormal outliers. some outlier detection methods distinguish the outliers by assigning each input data to object a label as either "normal" or "outlier" .while some other methods use the score measure as the factor to decide whether the object is an outlier. Based upon the application consistency and its data type the outlier detection method is chosen.
- **Application-specific outlier detection:** The relationship model is dependent on the type of application and it describes the normal data objects characteristics. Different applications require different types of data as input and require various modeling and analysis algorithms. Example: In clinical data analysis, a small deviation of data values reflects the choice of an outlier. In contrast, in marketing analysis, a larger deviation of data values is needed to justify an outlier. Choosing the Outlier detection method depends on the application type. We need to find out the outliers from a vast variety of applications data so the data types of these data sets may vary. There is no unique outlier detection method for all the applications.
- **Handling the noise in outlier detection:** Noise is usually present in all the data sets. Noise is present in outliers also. But there is a misassumption that noise and outliers are the same. The noise makes the quality of the data set to be imperfect. Noise often occurs when the data is

The challenges of outlier detection methods in data mining are listed below.

- **Modeling normal outliers effectively:** The quality of Outlier detection depends on the modeling of normal (which are not an outlier) objects. Often, building a model for finding the data normality is very challenging and maybe impossible because it is hard to determine all the behavioral properties of the normal objects. It is difficult to predict the border between normal outliers and abnormal outliers. some outlier detection methods distinguish the outliers by assigning each input data to object a label as either "normal" or "outlier" .while some other methods use the score measure as the factor to decide whether the object is an outlier. Based upon the application consistency and its data type the outlier detection method is chosen.
- **Application-specific outlier detection:** The relationship model is dependent on the type of application and it describes the normal data objects characteristics. Different applications require different types of data as input and require various modeling and analysis algorithms. Example: In clinical data analysis, a small deviation of data values reflects the choice of an outlier. In contrast, in marketing analysis, a larger deviation of data values is needed to justify an outlier. Choosing the Outlier detection method depends on the application type. We need to find out the outliers from a vast variety of applications data so the data types of these data sets may vary. There is no unique outlier detection method for all the applications.
- **Handling the noise in outlier detection:** Noise is usually present in all the data sets. Noise is present in outliers also. But there is a misassumption that noise and outliers are the same. The noise makes the quality of the data set to be imperfect. Noise often occurs when the data is collected from many resources and applications. Noise in the data sets is caused due to the duplicate tuples, missing values, and deviation of data attributes. Noise in the data sets makes the data-poor and it becomes a huge challenge to outlier detection. If noise is present in the data then it becomes difficult to retrieve the normal objects and separate the outliers from the data sets. Missing values may hide outliers and reduce the chance of detection of outliers.
- **Understandability:** In some cases, a client requires the condition of why a particular object has become an outlier as it may be useful for the process of applications. There must be a specific conditional criterion and justification to distinguish the normal objects from the outliers. And that justification must be well formulated. and understandable. Example: It is clear to understand the proximity outlier detection as the normal objects have nearly the proximity measures where as the outliers differ a large in their proximity measure.

Other Challenges Include

- **Heterogeneity of data:** It is difficult to create a universal outlier identification technique that works with all data types since datasets frequently comprise a variety of data kinds (text, numeric, categorical, etc.).
- **Scalability:** Scalable outlier identification systems are necessary to effectively manage huge datasets. Using large datasets may cause some procedures to become unmanageable due to the increasing computing needs.
- **Dimensions:** The curse of dimensionality makes it more difficult to separate outliers from regular patterns in high-dimensional datasets, which presents difficulties. In high-dimensional spaces, the effectiveness of traditional outlier detection techniques may diminish.
- **Variability and Noise:** It can be challenging to discern between true outliers and random oscillations in data due to noise and variability. To manage noisy data, robust outlier detection techniques are required.
- **Algorithm Sensitivity:** It might be difficult to determine how sensitively outcomes are to parameter changes in outlier identification methods, which frequently require parameter adjustment. Subjective outcomes could arise from parameter setting without a standard method.
- **Managing Distributions with Many Modes:** Differentiating between true outliers and cases from other modes can be challenging when dealing with datasets that have several modes or clusters. This is because traditional approaches may not be able to make this distinction.

Conclusion

The difficulties in detecting outliers highlight the necessity of a sophisticated and context-sensitive strategy. Ongoing research and innovation are essential for overcoming these obstacles and improving the durability of outlier detection techniques in the field of data mining as technology advances and datasets become more diverse. Despite the difficulties, solving these problems could lead to insightful discoveries and raise the general dependability of data-driven decision-making procedures.

You

Outlier Analysis: Types of outliers, Challenges of Outlier Detection, Outlier Detection Methods. Explain each topic in detail

ChatGPT

Outlier Analysis: Types of Outliers, Challenges of Outlier Detection, Outlier Detection Methods

Types of Outliers

Outliers are data points that deviate significantly from other observations in a dataset. They can be categorized into several types based on their nature and the context in which they occur:

1. **Global Outliers (Point Anomalies):**
   - **Definition:** These are individual data points that are significantly different from the rest of the data.
   - **Example:** In a dataset of people's ages, a data point of 150 years would be a global outlier.
2. **Contextual Outliers (Conditional Anomalies):**
   - **Definition:** These are data points that are considered outliers in a specific context or condition but may be normal in another context.
   - **Example:** A temperature of 30°C is normal in summer but would be a contextual outlier in winter.
3. **Collective Outliers:**
   - **Definition:** A collection of data points that collectively deviate significantly from the overall data pattern, even though individual points might not be outliers.
   - **Example:** A sudden drop in stock prices across multiple companies in a specific sector could be a collective outlier indicating a market crash.
4. **Novelty Outliers:**
   - **Definition:** These are new, previously unseen data points that may indicate a change in the underlying process.
   - **Example:** An unexpected spike in website traffic due to a viral post.

Challenges of Outlier Detection

Detecting outliers is fraught with several challenges:

1. **High Dimensionality:**
   - **Issue:** As the number of dimensions (features) increases, the concept of proximity or distance becomes less meaningful.
   - **Impact:** It becomes harder to distinguish between normal data points and outliers.
2. **Small Sample Size of Outliers:**
   - **Issue:** Outliers are rare by definition, making them hard to detect, especially in large datasets.
   - **Impact:** Methods must be sensitive enough to detect rare occurrences without being overwhelmed by normal data.
3. **Noise in Data:**
   - **Issue:** Data can be noisy, and distinguishing noise from true outliers can be difficult.
   - **Impact:** Noise can either mask outliers or be mistaken for outliers, leading to false positives or negatives.
4. **Varying Data Distribution:**
   - **Issue:** Outliers can appear in different forms depending on the data distribution.

4. **Varying Data Distribution:**
   - **Issue:** Outliers can appear in different forms depending on the data distribution.
   - **Impact:** Methods must be adaptable to different distributions to effectively detect outliers.
5. **Scalability:**
   - **Issue:** Many outlier detection methods do not scale well to large datasets.
   - **Impact:** Efficient and scalable algorithms are needed to handle big data.
6. **Interpretability:**
   - **Issue:** Once outliers are detected, understanding why they are outliers is crucial.
   - **Impact:** Methods should provide insights into the nature of outliers for meaningful analysis.

Outlier Detection Methods

There are several methods for detecting outliers, each suited to different types of data and applications:

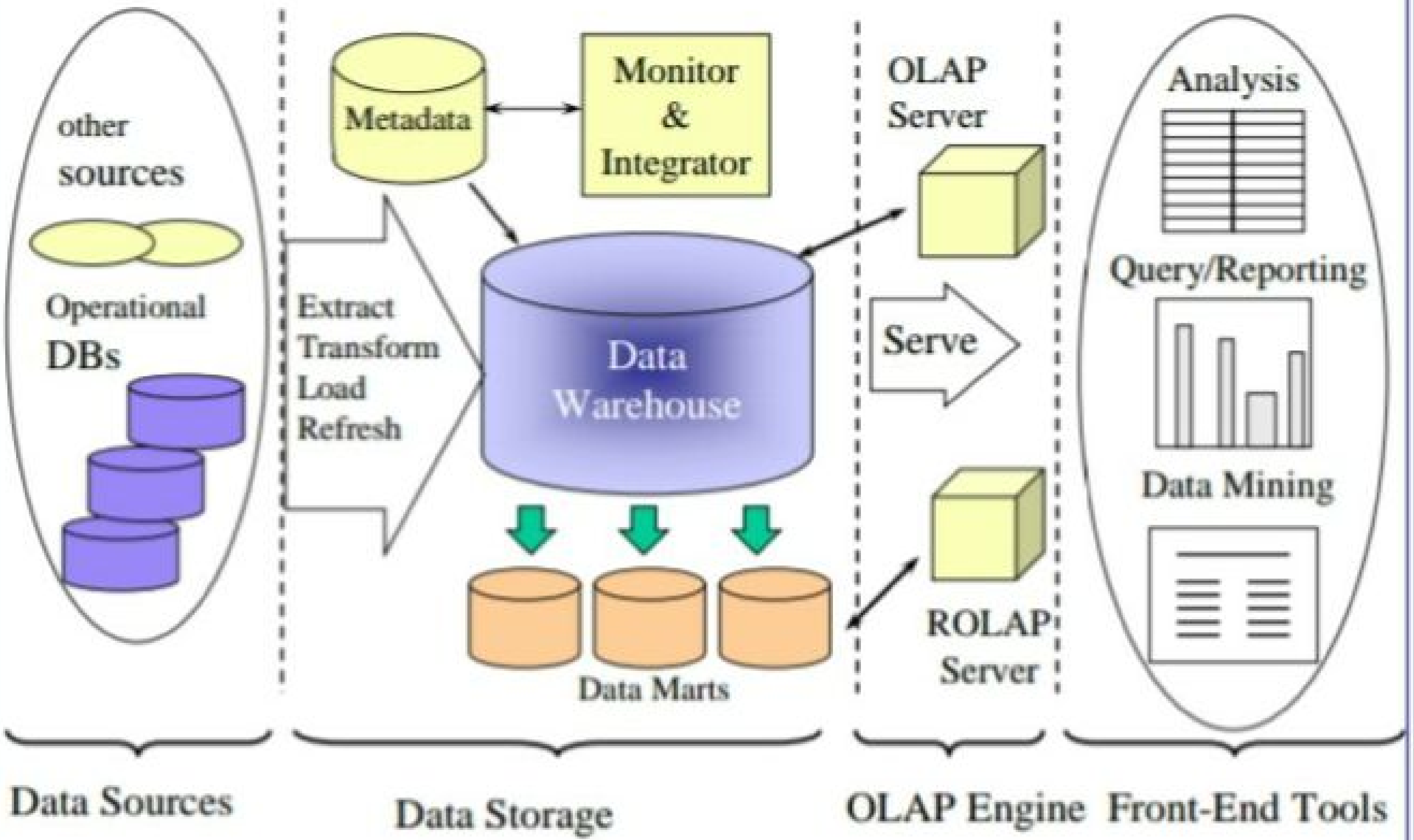
1. **Statistical Methods:**
   - **Z-Score:** Measures how many standard deviations a data point is from the mean.
     - **Usage:** Effective for normally distributed data.
     - **Example:** Data points with a Z-score greater than 3 or less than -3 are often considered outliers.
   - **IQR (Interquartile Range):** Uses the range between the first quartile (Q1) and third quartile (Q3).
     - **Usage:** Effective for data without a known distribution.
     - **Example:** Data points outside the range [Q1 - 1.5IQR, Q3 + 1.5IQR] are considered outliers.
2. **Distance-Based Methods:**
   - **Euclidean Distance:** Calculates the distance between points in multi-dimensional space.
     - **Usage:** Effective for small to medium-sized datasets with low dimensionality.
     - **Example:** Points that have a large distance from their nearest neighbors are considered outliers.
   - **Mahalanobis Distance:** Takes into account the correlations between variables.
     - **Usage:** Effective for datasets with correlated features.
     - **Example:** Points with a high Mahalanobis distance from the mean are considered outliers.
3. **Density-Based Methods:**
   - **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):**
     - **Usage:** Identifies clusters of points and labels points in low-density regions as outliers.
     - **Example:** Points not belonging to any cluster are considered outliers.
   - **LOF (Local Outlier Factor):**
     - **Usage:** Measures the local density deviation of a data point with respect to its neighbors.
     - **Example:** Points with a low density compared to their neighbors are considered outliers.
4. **Model-Based Methods:**
   - **Isolation Forest:**
     - **Usage:** Constructs trees where outliers are more isolated.
     - **Example:** Points that require fewer splits to be isolated are considered outliers.
   - **Autoencoders (in Neural Networks):**

2. **Distance-Based Methods:**
   - **Euclidean Distance:** Calculates the distance between points in multi-dimensional space.
     - **Usage:** Effective for small to medium-sized datasets with low dimensionality.
     - **Example:** Points that have a large distance from their nearest neighbors are considered outliers.
   - **Mahalanobis Distance:** Takes into account the correlations between variables.
     - **Usage:** Effective for datasets with correlated features.
     - **Example:** Points with a high Mahalanobis distance from the mean are considered outliers.
3. **Density-Based Methods:**
   - **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):**
     - **Usage:** Identifies clusters of points and labels points in low-density regions as outliers.
     - **Example:** Points not belonging to any cluster are considered outliers.
   - **LOF (Local Outlier Factor):**
     - **Usage:** Measures the local density deviation of a data point with respect to its neighbors.
     - **Example:** Points with a low density compared to their neighbors are considered outliers.
4. **Model-Based Methods:**
   - **Isolation Forest:**
     - **Usage:** Constructs trees where outliers are more isolated.
     - **Example:** Points that require fewer splits to be isolated are considered outliers.
   - **Autoencoders (in Neural Networks):**
     - **Usage:** Uses a neural network to reconstruct input data; high reconstruction error indicates outliers.
     - **Example:** Points with high reconstruction error are considered outliers.
5. **Clustering Methods:**
   - **K-Means Clustering:**
     - **Usage:** Points that do not fit well into any cluster can be considered outliers.
     - **Example:** Points with high distances to cluster centroids are considered outliers.
   - **Hierarchical Clustering:**
     - **Usage:** Points that form small, distinct clusters can be considered outliers.
     - **Example:** Single-member clusters are often outliers.
6. **Ensemble Methods:**
   - **Combination of Multiple Methods:**
     - **Usage:** Combines several methods to improve robustness and accuracy.
     - **Example:** An outlier identified by multiple methods is more likely to be a true outlier.

Conclusion

Outlier analysis is a critical aspect of data analysis, providing insights into anomalies that could indicate errors, fraud, or novel phenomena. Understanding the types of outliers, the challenges involved in detecting them, and the variety of methods available is essential for effective outlier detection and analysis. Each method has its strengths and weaknesses, and often a combination of methods is used to achieve the best results.

Three-Tier Architecture



Difference between OLAP and OLTP

Category	OLAP (Online Analytical Processing)	OLTP (Online Transaction Processing)
Definition	It is well-known as an online database query management system.	It is well-known as an online database modifying system.
Data source	Consists of historical data from various Databases.	Consists of only operational current data.
Method used	It makes use of a data warehouse.	It makes use of a standard database management system (DBMS).
Application	It is subject-oriented. Used for Data Mining, Analytics, Decisions making, etc.	It is application-oriented. Used for business tasks.
Normalized	In an OLAP database, tables are not normalized.	In an OLTP database, tables are normalized (3NF).
Usage of data	The data is used in planning, problem-solving, and decision-making.	The data is used to perform day-to-day fundamental operations.
Task	It provides a multi-dimensional view of different business tasks.	It reveals a snapshot of present business tasks.
Purpose	It serves the purpose to extract information for analysis and decision-making.	It serves the purpose to Insert, Update, and Delete information from the database.
Volume of data	A large amount of data is stored typically in TB, PB	The size of the data is relatively small as the historical data is archived in MB, and GB.
Queries	Relatively slow as the amount of data involved is large. Queries may take hours.	Very Fast as the queries operate on 5% of the data.
Update	The OLAP database is not often updated. As a result, data integrity is unaffected.	The data integrity constraint must be maintained in an OLTP database.
Backup and Recovery	It only needs backup from time to time as compared to OLTP.	The backup and recovery process is maintained rigorously
Processing time	The processing of complex queries can take a lengthy time.	It is comparatively fast in processing because of simple and straightforward queries.
Types of users	This data is generally managed by CEO, MD, and GM.	This data is managed by clerksForex and managers.
Operations	Only read and rarely write operations.	Both read and write operations.
Updates	With lengthy, scheduled batch operations, data is refreshed on a regular basis.	The user initiates data updates, which are brief and quick.
Nature of audience	The process is focused on the customer.	The process is focused on the market.
Database Design	Design with a focus on the subject.	Design that is focused on the application.
Productivity	Improves the efficiency of business analysts.	Enhances the user's productivity.

Page No ____
Date

DWDM

⇒ DWDM Definition

⇒ DWDM Advantages / Disadvantages

- Disadvantages

  Privacy Issue - Customer information get stolen from busin
  Security Issue - Employee " " " " " "
  Misuse of Information - Unethical use of information by businesses to take benefit of vulnerable people.

⇒ DWDM Application

- Retail Industry → Product pricing
  → Purchase trends
  → Customer Segmentation
  → Effectiveness of Advertisements

- Finance & Banking → Predict Loan Payment
  → Identify Fraudulent activities
  → Analyze Customer Credit Policy
  → Classify & cluster customers for marketing

- Healthcare & Finance → To survive in the market insurance company requires information about customers & competitors With the help of data mining fraudulent behaviour can be detected & prevented to a greater extent

- Telecommunication → Identify Telecommunication involved
  → " call patterns
  → Catch Fraudulent Activities
  → Improve the quality of services
  → Identification of products & Services which provides the highest profits

Page No.
Date

- Higher Education → EDM (Educational Data Mining) is used
  → Predicting Students Future
  → Learning Behaviour
  → Study Patterns
  → Better teaching Techniques can be developed for better understanding.

- Marketing & Sales → Identification of products & Services which provides the highest profits
  → Customers Buying Patterns
  → Cross Market Analysis

- E-commerce - Data Mining is use to give best seller options to customer with its predicting analysis algorithm
  - Collects information related to
  - How many customers viewed a product
  - " " " purchased that product
  - " " " reviewed " "
  Ex Amazon

- Research Analysis → Biological Data Analysis
  → Fluid Dynamics
  → Chemical engineering
  → Ecosystem Modelling & Weather Forec

- Data Mining for crime industry

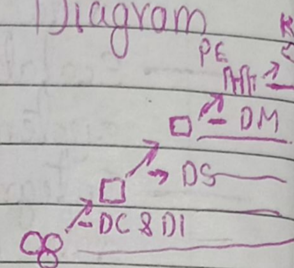
Page No.
Date

⇒ Data Mining Trends
— Application exploration - Development of application specific data mining systems
— Scalable data mining methods - Guide data mining systems in their search for interesting patterns
— Integration of data mining with database systems, data warehouse systems & web database systems
— Privacy Protection & information security in data mining
— Standardization of data mining language - A standard will facilitate systematic development, ~~improve~~ & promote the education & use of data mining systems in industry & society.
— New methods for mining complex types of data - More research is required towards the integration of data mining methods with existing data analysis techniques for the complex types of data.

⇒ Data Mining Sources / What kind of data can be mined
- Spatial Databases - Stores Geographical information like maps
- Flat Files - Text files - Data can be transaction, time series etc
- Relational Databases - Stores information in the form of a Table
- Transactional Databases - Stores transactional data with time series
- Multimedia Databases - Stores information in the form of video, image, audio & Text. Multimedia databases requires computer vision, computer graphics, image interpretation & NLP techniques.
- Data Warehouse - Data Warhousing is a collection data & information collected from multiple heterogeneous sources into one comprehensive database. And is often intended to be used as a whole under one unified schema.
- WWW (World Wide Web)

⟹ KDD Process - Data Cleaning
- Data Integration
- Data Selection
- Data Transformation
- Data Mining -
- Pattern Evaluation
- Knowledge Representation

+ Diagram



- Data cleaning
  - Noise & irrelevant data is removed
  - Outcome ~~of~~ would be dependent on the quality
  - Duplicate records are removed
  - Missing values are filled
  - Unnessasary data fields are removed

- Data Integration - It involves merging of data from differ data sources in order to form a data worehouse.
- Data Selection - Most Relevant data required for data minin is selected.
- Data Transformation - Data is transformed into appropriate form making it ready for data mining step
- Data Mining - Various data Mining techniques is applied to ~~detect~~ discover patterns. ~~Techniques such as~~ Data Mining algorithms such as classification, clustering, regression or association rule mining are applied to extract useful information.
- ~~Data~~ Pattern Evaluation - Once the patterns are discov they need to be evaluated to determine their quality, significance & usefulness
- Knowledge Representation - The Knowledge discoved is consalidated & represented to the user in a simple & easy to understand format.

Page No
Date

⇒ Types of Attributes

- Nominal - It means relating to names. The values of Nominal attributes are symbols or name of things. Each value represent some kind of category. And are also called categorical data. Not comparable or rankable.

- Ordinal - It is an attribute with possible values that have meaningful order or ranking among them. Ex Grade, Ranking, height (tall, medium, short).

- Binary - Takes two values, Yes or No, True or False
  - Symmetric - Both values are equally important (Gender)
  - Assymetric - " " " not equally important (Cancer)

- Interval - It comes with their ability dip below 0. Interval scale hold no true 0 value & can represent values below 0. Ex Temperature

- Ratio - Height & Weight are ratio. Ex Kelvin, length, time, counts.

Discrete	Continous
- It has only a finite set of values. Ex zip code or set of words	It has real numbers as attribute values. Ex height
- Binary is an subset of Discrete	

Page No.
Date

⇒ Major Issues in Data Minning
⇒ Associal Rule Mining - Association Rule mining finds interesting associations & relationships among large sets of data items. This rule shows how frequently a item set occurs in a transaction. A typical example is Market based analysis.

Market Based analysis is one of the key techniques used by large relations to show association b/w items. It allow retailers to identifies relationships b/w the items that people buy ~~f~~ frequently togeth

Example
Milk → Bread (Support = 2%, Confidence = 60%)

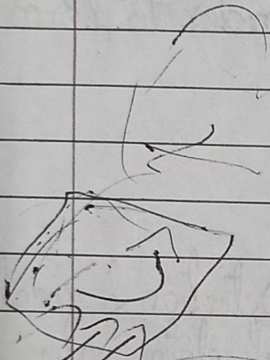
A support of 2% for association rule means that 2% of all the transactions show that milk & bread are purchased togther (support indicates how frequently item appears in the database). And 60% of confidence means 60% of all the customers who buy milk also bought bread.

⇒ Apriorit Numerical + FP Growth + Numerical

Data Preprocessing

⇒ Real databases are highly susceptible to noise & inconsistent data due to their typically huge size and heterogeneous sources. Low quality data will lead to low quality mining results. In order to improve the readability & speed of mining process data preprocessing is required

⇒ Data mining Tasks/steps
- Data cleaning
- Data integration
- Data Reduction or selection
- Data Transformation / Data Discretization.



Data Cleaning

Missing / Incomplete	Noisy	Inconsistent
eg Null or empty	eg Salary = -10 Age = -9	eg Age 50 DOB- 3/10/2000
Solution (Drop whole row or column)	Solution	Regression - Numerical prediction of data. It aims to predict the value of a dependent variable based on one or more independent variable.
- Ignore row or tuple	• Binning	
- Manually fill	- Smoothing by mean	
- Use Global Constant	- " " median	
- Use the attribute mean	- " " Boundary	
- Most Probable value	• Regression - Data can be smoothen by fitting it to a regression function	
Replace it with an avg (Mean & Median)	• Clustering (outliers)	
Replace it by frequency (Mode)		

- Clustering - Similar data items are grouped together. Dissimilar item are always outside the cluster and we tend to reject them. It is a ~~similar~~ technique in ML & data analysis that groups similar items together into cluster based on their characteristics or features

Page No
Date

It is an unsupervised learning method, which means the algorithm tries to find patterns in the data without any prior knowledge of what cluster should look like.

Data Integration

- It is a preprocessing method that involves merging of data from different data sources in order to form a data store ie data warehouse.

- Issues in Data Integration

Schema integration & Object Making	Redundancy Unwanted attributes should be removed	Detection & resolut[ion] of data value conflicts
Eg A Emp_no | Name | DOB B Emp_ID | Name | DOB During Integration Emp_no & Emp_ID can cause error so it needs to be resolved	Eg EmpID | name | DOB | Age Age is redundant data as we can calculate age with DOB	Eg A & B are compa[nies] A store price in Rs B store price in $ Correctly Modify th[e] data

Data Reduction

- Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same analytical results.

Types of Data Reduction

- Dimentionality Reduction OR Data compression - It represents the original data in the compressed or reduced form by applying data encoding or transformation

Loss Less	Lossy
If original data can be reconstructed from compressed data without lossing any info	If reconstructed is the approximation of compressed data

- Numerosity Reduction - It reduces the data volume by choosing alternative smallers forms of data representation

- Data cube aggregation - It is a process in which info is gathered & expressed in a summary form. Data in smaller volume or size.

Year Sales 2017

Year	Sales
H1	500
H2	300

Year Sales 2018

H1	600
H2	100
Year	sales

→

Year	Sales
2017	800
2018	700

Page No. ______
Date

Data Transformation

- A Function that maps the entire set of value of a given attribute to a new set of replacement values each old value can be identified with one of the new values.

- Methods

① ~~Smoothing~~
- ~~Binning~~
- ~~Regression~~
- ~~Clustering~~

② Aggregation Summarizing

③ Normalization
Converts data into the range of [0,1]. Done using Mean & standard deviation

- Min Max Normalization
- Z score Normalization
- Normalization by decimal scalling

④ Attribute Construction
New attributes constructed from the given ones for minning.

⑤ ~~Dimentionality reduction~~

Data Discretization

- Reduce the number of value for a given continous attribute by dividing the range of the attribute into intervals or Intervals labels can be used to replace actual data values. Such as numeric values for the attribute age by higher level concepts such as young, middle aged or senior.

Page No. ______
Date

Binning	Histogram	Clustering	Decision Tree	Correlation
Unsupervised	Unsupervised	Unsupervised	Supervised	Unsupervised
Top - down	Top - Down	Bottom down	Top down	Bottomup

Supervised

Regression

Classification

Binary classification

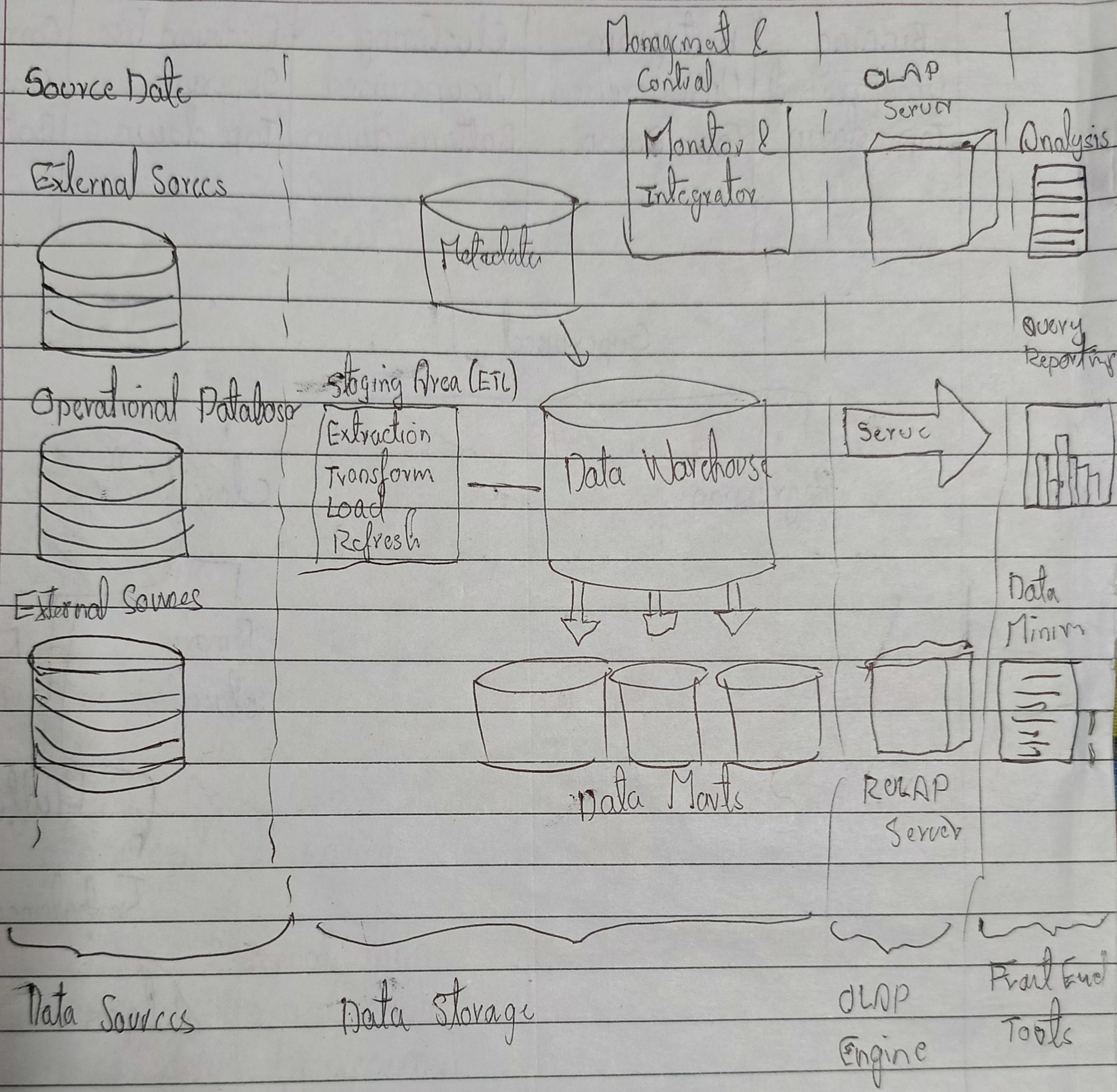
Multiclass Classification

Multilabel

Imbalanced Disease

Algos

categories



Source Date
External Sorces
Operational Database
External Sounes
Managmant &
Contiral
Monitor &
Integrator
Metadata
Staging Area (ETL)
Extraction
Transform
Load
Refresh
Data Warehouse
Data Marts
OLAP
Server
Serve
ROLAP
Server
Analysis
Query
Reporting
Data
Minim
Data Sources
Data Storage
OLAP
Engine
Frant End
Tools