DATA ANALYTICS

Sixth Semester B.Sc Computer Science

Study Material under the Syllabus of University of Kerala

2024

MUSLIM ASSOCIATION
COLLEGE OF ARTS & SCIENCE
Panavoor, Thiruvananthapuram

Prepared By
Prasanth B
Assistant Professor
Muslim Association College of Arts & Science

MUSLIM ASSOCIATION COLLEGE OF ARTS AND SCIENCE

Panavoor,Thiruvananthapuram,Kerala

(*Affiliated to the University of Kerala***)**

Department of Computer Science

CS1641 : DATA ANALYTICS

Name : ……

Candidate Code: ……

CS1641 : DATA ANALYTICS

SYLLABUS

MODULE I: - An Introduction to Data Analysis - Data Analysis, **Knowledge Domains of the Data Analyst** - Computer Science, Mathematics and Statistics, Machine Learning and Artificial Intelligence Professional Fields of Application. **Introduction to Big Data Analytics: -** Big Data Overview, State of the Practice in Analytics, Key Roles for the New Big Data Ecosystem. **Characteristics of Big Data-**Volume, Velocity, Variety, Veracity, Value. **Data Analytics Lifecycle: -** Data Analytics Lifecycle Overview, Discovery**,** Data Preparation, Model Planning, Model Building, Communicate Results, Operationalise , Case Study: Global Innovation Network and Analysis (GINA).

MODULE II: Advanced Analytical Theory and Methods: Clustering - Overview of Clustering, Kmeans,Additional Algorithms. **Advanced Analytical Theory and Methods: - Association Rules -** Overview, Apriori Algorithm, Evaluation of Candidate Rules, An Example: Transactions in a Grocery Store. **Advanced Analytical Theory and Methods:** Introduction to regression and classification.

MODULE III: Advanced Analytical Theory and Methods: Text Analysis - Text Analysis Steps, A Text Analysis Example, Collecting Raw Text, Representing Text, Term Frequency-Inverse Document Frequency (TFIDF), Categorizing Documents by Topics, Determining Sentiments, Gaining Insights.

MODULE IV: Advanced Analytics- Technology and Tools: MapReduce and Hadoop – Introduction to MapReduce and Apache Hadoop. The Hadoop Ecosystem – Pig, Hive, HBase, Mahout, NoSQL.**Advanced Analytics- Technology and Tools: In-Database Analytics: -** SQL Essentials – Joins, SetOperations. In-Database Text Analysis, **Data Privacy and Ethics: -** Privacy Landscape, Rights and Responsibility, Technologies.

MODULE I

An Introduction to Data Analysis - Data Analysis, **Knowledge Domains of the Data Analyst** - Computer Science, Mathematics and Statistics, Machine Learning and Artificial Intelligence Professional Fields of Application. **Introduction to Big Data Analytics: -** Big Data Overview, State of the Practice in Analytics, Key Roles for the New Big Data Ecosystem. **Characteristics of Big Data-**Volume, Velocity, Variety, Veracity, Value. **Data Analytics Lifecycle: -**Data Analytics Lifecycle Overview, Discovery**,** Data Preparation, Model Planning, Model Building, Communicate Results, Operationalise , Case Study: Global Innovation Network and Analysis (GINA).

Introduction to Data Analysis

- Before jumping into the term "**Data Analysis",** let's discuss the term **"Analysis"**.
- Analysis is a process of answering **"How?"** and **"Why?"**.
- For example, how was the growth of XYZ Company in the last quarter? Or why did the sales of XYZ Company drop last summer?
- So to answer those questions we take the data that we already have. Out of that, we filter out what we need.
- This filtered data is the final dataset of the larger chunk that we have already collected and that becomes the target of **data analysis**.

What is Data Analysis

- Data analysis is the process of examining, filtering, adapting, and modeling data to help solve problems. Data analysis helps determine what is and isn't working, so you can make the changes needed to achieve your business goals.
- A data analyst is a problem solver who prepares and analyzes data to provide organizations with insights that help them make better business decisions.
- Data analysts collect, organize, and analyze data sets to help companies or individuals make sense of information and drive smarter decision-making.
- Data analysis is a subset of data analytics.
- It is the technique of observing, transforming, cleaning, and modeling raw facts and figures with the purpose of developing beneficial information and acquiring profitable conclusions.

What is Data Analytics

- Analytics is a technique of converting raw facts and figures into some particular actions by analyzing those raw data evaluations and perceptions in the context of organizational problem-solving and also with the decision making.
- Analytics is the discovery and conversation of significant patterns in data.
- The aim of Data Analytics is to get actionable insights ensuing in smarter selections and higher commercial enterprise outcomes.

Difference between Data Analytics and Data Analysis :

S.No.	Data Analytics	Data Analysis
1.	It is described as a traditional form or generic form of analytics.	It is described as a particularized form of analytics.
2.	It includes several stages like the collection of data and then the inspection of business data is done.	To process data, firstly raw data is defined in a meaningful manner, then data cleaning and conversion are done to get meaningful information from raw data.
3.	It supports decision making by analyzing enterprise data.	It analyzes the data by focusing on insights into business data.
4.	It uses various tools to process data such as Tableau, Python, Excel, etc.	It uses different tools to analyze data such as Rapid Miner, Open Refine, Node XL, KNIME, etc.
5.	Descriptive analysis cannot be performed on this.	A Descriptive analysis can be performed on this.
6.	One can find anonymous relations with the help of this.	One cannot find anonymous relations with the help of this.
7.	It does not deal with inferential analysis.	It supports inferential analysis.

Knowledge Domains of the Data Analyst

1.Computer Science

- Computer science uses data analytics to help individuals and organizations make sense of data.
- Data analysts use data to create stories by modulating data requirements for collection, processing, cleaning, and exploratory analysis.
- In the context of computer science, data analysis involves modulating data requirements appropriate for data collection, processing, cleaning, and exploratory analysis.
- Basically, a data analyst takes a data source, and uses the data to create a story.
- These stories are visualizations that effectively convey information.

Here are some ways computer science uses data analytics:

1. Data collection: Computer science helps with understanding and working with aspects of big data.
2. Data pre-processing: Computer science helps with cleaning and SQL.
3. Analysis: Computer science helps with analysis, including EDA.
4. Insights: Computer science helps with machine learning and deep learning.
5. Visual reports: Computer science helps with visualizations

2.Mathematics and Statistics

- Math and Statistics for Data Science are essential because these disciples form the basic foundation of all the Machine Learning Algorithms.
- In fact, Mathematics is behind everything around us, from shapes, patterns and colors, to the count of petals in a flower.
- Mathematics is embedded in each and every aspect of our lives.
- To become a successful Data Scientist you must know your basics.
- Math and Stats are the building blocks of Machine Learning algorithms.
- It is important to know the techniques behind various Machine Learning algorithms in order to know how and when to use them.

- Statistics is a Mathematical Science pertaining to data collection, analysis, interpretation and presentation.
- Statistics can be used to derive meaningful insights from data by performing mathematical computations on it.
- Several Statistical functions, principles and algorithms are implemented to analyse raw data, build a Statistical Model and infer or predict the result.
- Types Of Analysis -Math And Statistics For Data Science are as follows

1. **Quantitative Analysis:** Quantitative Analysis or the Statistical Analysis is the science of collecting and interpreting data with numbers and graphs to identify patterns and trends.
2. **Qualitative Analysis:** Qualitative or Non-Statistical Analysis gives generic information and uses text, sound and other forms of media to do so.

3.Machine Learning

- Machine Learning provides techniques to extract data and then appends various methods to learn from the collected data and then with the help of some well-defined algorithms to be able to predict future trends from the data.
- Machine Learning or traditional machine learning had its core revolving around spotting patterns and then grasp the hidden insights of the available data.
- For any business, industry, and organization to run data as a primary record or lifeblood of it, and along with evolution, there is also a rise in demand and importance. This aspect is why data engineers and data scientists need machine learning.
- With the help of this technology, you can analyze a large amount of data and calculate risk factors in no time.

  Example
- Google is the quintessential example for machine learning as GOOGLE records the number of searches you have made and then suggests you similar searches when you google something in the future. Similarly, AMAZON recommends your products based

on your previous searches and so does NETFLIX, based on the TV show or Movies that you watched, you get a similar type of suggestions.

4. Artificial intelligence

- Artificial intelligence (AI) data analysis uses AI techniques and data science to improve the processes of cleaning, inspecting, and modeling structured and unstructured data.
- The goal is to uncover valuable information to support drawing conclusions and making decisions.
- AI can identify patterns and correlations that are not obvious to the human eye, making it a more effective tool for data analysis.
- AI can also scan datasets for errors, inconsistencies, and anomalies and immediately rectify them.
- AI cognitive analysis involves the use of AI to simulate human thought processes in a computerized model. This allows for quick and accurate analysis of large amounts of unstructured data.
- Some examples of AI tool in Data Analytics
    - Ajelix: A top AI tool for Excel.
    - Arcwise AI: An advanced AI-powered Excel tool.
    - Sheet+: A top AI formula generator Excel tool.

What is Big Data

- Data which are very large in size is called Big Data. Normally we work on data of size MB(WordDoc ,Excel) or maximum GB(Movies, Codes) but data in Peta bytes i.e. 10^15 byte size is called Big Data.
- Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low latency.
- Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources, and in different sizes from terabytes to zetta bytes.
- Examples of Big Data are-Facebook , New York Stock Exchange
- Big data has one or more of the following characteristics:
    1. High volume
    2. High velocity
    3. High variety.

Sources of Big Data

- **Social networking sites:** Facebook, Google, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.
- **E-commerce site:** Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.
- **Weather Station:** All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.
- **Telecom company:** Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.
- **Share Market:** Stock exchange across the world generates huge amount of data through its daily transaction.

Types of Big Data

BigData' could be found in three forms:

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

1.Structered Big Data

- Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.
- Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it.
- However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes.
- Example

2.Unstructered Big Data

Sl No	Name	Class	Mark	Place
1	Abhi	MCA	98	Trivandrum
2	Kiram	BSc	92	Kollam
3	Seena	MBA	98	Kollam

- Any data with unknown form or the structure is classified as unstructured data.
- In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it.
- A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc.
- Now day organizations have wealth of data available with them but unfortunately, they don't know how to derive value out of it since this data is in its raw form or unstructured format.
- **Examples:** The output returned by 'Google Search'

3.Semi Structured Big Data

- Semi-structured data can contain both the forms of data.
- We can see semi-structured data as a structured in form but it is actually not defined
- Example of semi-structured data is a data represented in an XML file.

3V's of Big Data

1. **Velocity:** The data is increasing at a very fast rate. It is estimated that the volume of data will double in every 2 years.
2. **Variety:** Now a days data are not stored in rows and column. Data is structured as well as unstructured. Log file, CCTV footage is unstructured data. Data which can be saved in tables are structured data like the transaction data of the bank.
3. **Volume:** The amount of data which we deal with is of very large size of Peta bytes.

State of the Practice of Analytics

- The State of the Practice of Analytics refers to the current state and ongoing evolution of the use of data, statistical analysis, and other quantitative methods to derive insights and inform decision-making.
- It encompasses the various applications of analytics across industries and domains, as well as the key trends and future directions in the field.
- The State of the Practice of Analytics is constantly evolving, driven by advances in technology, changing business needs, and evolving customer expectations.

The main Analytics are

1.Applications of Analytics

Analytics has a wide range of applications(software) across various industries and domains. Some of the most common applications of analytics include:

2.**Business analytics:** This involves the use of data to gain insights into business operations, such as sales, marketing, and finance.

For example, a retailer can use analytics to identify trends in customer behavior, optimize pricing strategies, and forecast demand for products.

3.Healthcare analytics: This involves the use of data to improve patient outcomes, reduce costs, and optimize healthcare delivery.

For example, a hospital can use analytics to identify high-risk patients, predict readmissions, and optimize staffing levels.

4.Fraud detection and prevention: This involves the use of analytics to identify and prevent fraudulent activities, such as credit card fraud and insurance fraud.

For example, a financial institution can use analytics to detect patterns of suspicious activity and flag potentially fraudulent transactions.

5.Social media analytics: This involves the use of data from social media platforms to understand customer behavior, sentiment, and engagement.

For example, a company can use analytics to track brand mentions, analyze customer feedback, and identify influencers who can help promote their products or services.

6.Predictive maintenance: This involves the use of analytics to predict when equipment or machinery is likely to fail, allowing for proactive maintenance and reducing downtime.

For example, a manufacturer can use analytics to monitor equipment performance, detect anomalies, and schedule maintenance before a breakdown occurs.

Key Trends in Analytics:

The state of the practice of analytics is constantly evolving, driven by advances in technology, changing customer expectations, and evolving business needs. Some of the key trends in analytics include:

1.Big data: The growth of data volumes, velocity, and variety is driving the need for new tools and technologies to process and analyze large datasets.

2.Artificial intelligence and machine learning: These technologies are enabling organizations to derive insights from data in real-time, automate decision-making, and enhance customer experiences.

3.**Cloud computing:** The cloud is increasingly becoming the preferred platform for analytics, offering scalability, flexibility, and cost-effectiveness.

4.Data visualization: As data becomes more complex, data visualization is becoming an increasingly important tool for communicating insights and findings to non-technical stakeholders.

Key Roles for the New Big Data Ecosystem

- The big data ecosystem refers to the complex network of technologies, tools, processes, and people involved in the collection, storage, processing, and analysis of large and diverse sets of data.
- The new big data ecosystem involves various components that play key roles in handling, processing, and deriving insights from vast amounts of data.

Some key roles within this ecosystem include:

1.Data Collection Tools and Technologies: These tools gather data from various sources, such as sensors, IoT devices, social media, websites, etc. They ensure the raw data is efficiently acquired and transmitted for further processing.

2.Data Storage Solutions: Components like data lakes, data warehouses, and distributed file systems store massive volumes of structured and unstructured data. These need to efficiently manage and organize data for quick retrieval and analysis.

3.Data Processing Frameworks: Technologies like Hadoop, Spark, and Flink handle the processing of large datasets, enabling parallel processing, real-time data streaming, and batch processing.

4.Data Cleaning and Preprocessing: This role involves tools and methodologies that clean, filter, and preprocess raw data to enhance its quality before analysis. Techniques like data normalization, deduplication, and outlier detection fall into this category.

5.Analytics and Visualization Tools: These tools help in analyzing and interpreting data. They include machine learning algorithms, statistical models, and visualization platforms that make complex data more understandable and actionable.

6.Data Governance and Security: Managing data access, ensuring compliance with regulations (like GDPR), maintaining data quality, and implementing security measures to protect sensitive information are crucial roles in the big data ecosystem.

7.Data Scientists and Analysts: Skilled professionals who interpret data, build models, and derive insights from the information gathered. They play a critical role in making sense of the data and turning it into actionable strategies.

Characteristics of Big Data

Big data is characterized by what's often referred to as the "three Vs" – Volume, Velocity, and Variety. Additionally, two more Vs have been added in some discussions, focusing on Veracity and Value. Here are the key characteristics:

1.Volume:

- Big data involves vast amounts of data generated from various sources, including business transactions, social media, sensors, and more.
- The volume refers to the sheer size of data, often ranging from terabytes to exabytes and beyond.

2.Velocity:

- Data is generated at an incredibly high speed.
- This refers to the rate at which data is produced, collected, processed, and analyzed in real-time or near real-time.
- For instance, streaming data from sensors, social media feeds, or financial transactions requires rapid processing to extract actionable insights.

3.Variety:

- Big data comes in different formats and types.

- It includes structured data (like databases), unstructured data (such as text, images, videos), and semi-structured data (like XML or JSON files).
- Managing and analyzing this diverse range of data types is a challenge in big data analytics.

4.Veracity:

- This aspect emphasizes the uncertainty or trustworthiness of available data.
- With large volumes of data coming from various sources, ensuring data quality, accuracy, and reliability becomes crucial.
- Veracity refers to the reliability of the data and the assurance that it is trustworthy for analysis and decision-making.

5.Value:

- The ultimate goal of big data analysis is to extract value from the data.
- Finding meaningful insights, making informed decisions, improving efficiency, identifying trends, and discovering new opportunities are some of the ways in which value is derived from big data analytics.

Data Analytics Lifecycle

The Data Analytics Lifecycle refers to the step-by-step process followed in the field of data analytics to derive insights and valuable information from data.

This lifecycle typically includes several stages:

1.Discovery

- This stage is critical in laying the foundation for subsequent analysis and decision-making.
- Discovery is foundational as it shapes subsequent steps in the analytics lifecycle, including model selection, feature engineering, and hypothesis testing.

- The process begins with a clear understanding of the business goals or problems that need to be addressed through data analysis.

2.Data Preparation

- Data preparation is a crucial phase in the data analytics lifecycle where raw data is cleaned, transformed, and organized to make it suitable for analysis.
- This stage is vital because raw data is often messy, inconsistent, and may contain errors, missing values, or irrelevant information.
- Data preparation ensures that the data is in a usable format for analysis and modeling.
- The main activity in data preparation phase is
- **Data Cleaning:** This involves handling missing values, removing duplicates, correcting errors, and dealing with inconsistencies in the dataset. Techniques like imputation (filling missing values with estimated ones) or deletion of irrelevant or redundant data fall under data cleaning.

3.Model Planning

- In this phase overall strategy for building analytical models is devised.
- This phase involves determining the objectives, selecting appropriate techniques, and outlining the approach for creating models that will best address the business problems or goals.
- Based on the insights gained, analysts or data scientists select appropriate algorithms and build models to extract further insights or make predictions.
- This stage involves machine learning, statistical modeling, or other analytical methods.

The main aim of this phase is

1. Understanding Business Objectives
2. Defining Success Criteria
3. Data Understanding
4. Model Selection
5. Prototyping and Experimentation
6. Risk Assessment

4.Model Building

- In this phase the chosen analytical models are developed and trained using the prepared dataset.
- This stage involves implementing the chosen algorithms, tuning parameters, and creating predictive or descriptive models to extract insights or make predictions.

The following operations are performed in this phase

1. **Data Splitting:** Dividing the dataset into subsets for training, validation, and testing.
2. **Algorithm Implementation:** Implementing the selected modeling algorithms or techniques based on the defined model plan.
3. **Model Training:** Using the training dataset to train the model on historical data.
4. **Validation:** Assessing the model's performance using the validation dataset.
5. **Documentation and Reporting:** Documenting the entire model-building process

5.Communicate Results

- Effective communication of results is crucial to ensure that the insights derived from data analysis are understood, accepted, and utilized for informed decision-making within the organization.
- It involves presenting insights, findings, and recommendations derived from the analysis to stakeholders in a clear, understandable, and actionable manner.

The main activities are

1. **Understand the Audience:** Tailor the communication to the audience's level of technical expertise and their specific needs.
2. **Summarize Key Findings:** Begin with a concise summary of the most critical insights and findings.
3. **Use Visualizations:** Utilize charts, graphs, and visual aids to present complex information in an easily understandable format.
4. **Provide Context:** Explain the context behind the data, methodologies used, and any assumptions made during the analysis

5. **Tell a Story:** Structure the presentation in a narrative format, guiding stakeholders through the analysis process step by step.

6.Operationalize

- " Operationalize " in the context of the data analytics lifecycle refers to the process of implementing the insights, models, or recommendations derived from data analysis into operational systems or workflows.
- It involves translating analytical findings into actions that impact business operations or decision-making processes.

The main activities are

1. **Deployment of Models:** After developing and testing analytical models, the operationalization phase involves deploying these models into production environments.
2. **Automation of Processes:** Implementing automated systems or workflows based on data-driven insights.
3. **Integration with Existing Systems:** Ensuring seamless integration of analytics results into existing business systems
4. **Monitoring and Maintenance:** Continuously monitoring the implemented models or systems to ensure they perform as expected.

Case Study: Global Innovation Network and Analysis (GINA).

GINA is a multinational initiative aimed at fostering innovation and collaboration among global entities across various industries. It operates as a network that collects, analyzes, and disseminates insights related to innovation trends, emerging technologies, and best practices.

Objectives:

1. ***Facilitating Innovation Exchange:*** *GINA aims to create a platform that enables the exchange of innovative ideas, technologies, and methodologies among global stakeholders.*

2. ***Insightful Analysis:*** *It focuses on analyzing global innovation trends, identifying disruptive technologies, and providing actionable insights to member organizations.*
3. ***Strategic Partnerships:*** *GINA seeks to establish strategic partnerships and collaborations between industry leaders, academia, and government bodies to drive innovation on a global scale.*
4. ***Promoting Best Practices:*** *It aims to identify and promote best practices in innovation, R&D, and technology adoption across diverse sectors.*

Key Components and Activities:

1. ***Data Collection and Aggregation:*** *GINA aggregates data from various sources including research publications, patent databases, industry reports, and innovation indices.*
2. ***Analytical Framework:*** *Utilizes advanced analytics, machine learning, and natural language processing to analyze the collected data, identifying trends, patterns, and emerging technologies.*
3. ***Insight Generation:*** *GINA generates actionable insights and reports on emerging technology domains, innovation hotspots, R&D investment trends, and market opportunities.*
4. ***Collaboration Platform:*** *Provides an online platform for member organizations to collaborate, share knowledge, and engage in joint innovation projects.*
5. ***Events and Workshops:*** *Organizes global events, workshops, and forums to facilitate networking, knowledge sharing, and ideation sessions among stakeholders*

************END OF MODULE 1************

MODULE II

Advanced Analytical Theory and Methods: Clustering - Overview of Clustering, Kmeans,Additional Algorithms. **Advanced Analytical Theory and Methods: - Association Rules -** Overview, Apriori Algorithm, Evaluation of Candidate Rules, An Example: Transactions in a Grocery Store. **Advanced Analytical Theory and Methods:** Introduction to regression and classification.

Advanced Analytical Theory and Methods: Clustering

Overview of Clustering

- Clustering is an essential tool in exploratory data analysis, providing insights into the structure of data when there's no predefined outcome to predict, making it a key component in understanding and interpreting complex datasets.
- Clustering is a technique used in data analysis and unsupervised machine learning. Its primary objective is to group similar objects or data points together in a way that items in the same group (or cluster) are more similar to each other than to those in other clusters.

The process of clustering involves

1. **Similarity Measurement**: Determining the similarity or dissimilarity between data points using metrics like Euclidean distance, cosine similarity, etc.
2. **Grouping Data:** Based on the similarity measures, the algorithm groups the data points into clusters. The number of clusters may be predefined or determined by the algorithm itself.
3. **Cluster Assignment:** Assigning each data point to the nearest cluster centroid or representative.
4. **Evaluation and Refinement:** Assessing the quality of clusters and refining them if necessary.

Uses of Clustering

1. **Pattern Recognition:** Identifying inherent patterns or structures within data.

2. **Market Segmentation:** Grouping customers or products based on similarities for targeted marketing or recommendation systems.
3. **Image Segmentation:** Partitioning an image into meaningful segments or regions.
4. **Anomaly Detection:** Detecting outliers or anomalies that don't conform to expected patterns within a dataset.

1.K-Means Algorithm

- K-Means is one of the most commonly used clustering algorithms.
- It's an iterative algorithm that partitions a dataset into K clusters where each data point belongs to the cluster with the nearest mean (centroid), serving as a prototype of the cluster.
- K-Means is computationally efficient and works well for medium to large-sized datasets.
- it's widely used in various applications like customer segmentation, image segmentation, document clustering, and more.
- However, for more complex cluster shapes or densities, other algorithms like DBSCAN or hierarchical clustering might be more appropriate.

K-Means algorithm:

1. **Initialization:** Choose K initial centroids randomly from the data points or place them strategically.
2. **Assignment Step**: Assign each data point to the nearest centroid. Typically, this is done by calculating distances (often Euclidean distance) between each point and each centroid and assigning the point to the nearest centroid.
3. **Update Step:** Recalculate the centroids of the newly formed clusters.
4. **Iteration:** Repeat the assignment and update steps until convergence. Convergence occurs when the centroids no longer change significantly or after a specified number of iterations.

2.DBSCAN Algorithm

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a powerful clustering algorithm that groups together points that are closely packed together, defining clusters as areas of high density separated by regions of low density.
- It's particularly effective in identifying clusters of arbitrary shapes and sizes and can handle noise (outliers) effectively without requiring the number of clusters to be predefined

The algorithm is based on two parameters:

1. **Epsilon (ε):** This parameter defines the radius within which to search for neighboring points.
2. **MinPts:** It specifies the minimum number of points within the ε radius to consider a point as a core point.

The steps involved in DBSCAN are:

1. **Core Point Identification:** For each data point, DBSCAN counts how many other points are within its ε neighborhood. If the number of points within ε is greater than or equal to MinPts, the point is marked as a core point.
2. **Expansion of Clusters:** Starting from a core point, the algorithm forms a cluster by including all reachable points (directly or indirectly) within ε distance. It iterates through these points, expanding the cluster until no more points can be added.
3. **Noise Identification:** Points that are not core points and do not belong to any cluster are considered noise/outliers.

Advanced Analytical Theory and Methods: - Association Rules

Overview of Association Rule

- Association rules in data analytics refer to patterns or relationships between variables or items in a dataset.
- These rules help uncover associations, correlations, or dependencies among different data points or variables, aiding in the understanding of data and decision-making processes.
- The primary purpose of association rule mining is to discover interesting relationships between variables/items in large datasets.
- Association rules are typically expressed in the form of "if-then" statements:
    - Antecedent (IF): The item or condition that precedes or is present.
    - Consequent (THEN): The item or condition that follows or is predicted to be present based on the antecedent.

Uses of Association Rule

1. **Market Basket Analysis:** Understanding the relationships between products purchased together in retail transactions. This helps in optimizing product placements, cross-selling, and targeted marketing strategies.
2. **Healthcare Analytics:** Discovering associations between symptoms and diseases to aid in diagnosis or understanding the risk factors associated with certain conditions.
3. **Web Usage Mining:** Analyzing user behavior on websites to recommend relevant content or understand browsing patterns.

4. **Fraud Detection:** Identifying associations between seemingly unrelated activities that might indicate fraudulent behavior.

Apriori Algorithm

- The Apriori algorithm is a fundamental algorithm used in association rule mining to discover frequent itemsets within transactional datasets and generate association rules based on these itemsets.
- It's widely used for market basket analysis, where the goal is to find relationships between items frequently purchased together.
- Apriori algorithm helps the customers to buy their products with ease and increases the sales performance of the particular store.

Components of Apriori algorithm

The given three components comprise the apriori algorithm.

1. Support
2. Confidence
3. Lift

Let's take an example to understand this concept.

We have already discussed above; you need a huge database containing a large no of transactions. Suppose you have 4000 customers transactions in a Big Bazar. You have to calculate the Support, Confidence, and Lift for two products, and you may say Biscuits and Chocolate. This is because customers frequently buy these two items together. Out of 4000 transactions, 400 contain Biscuits, whereas 600 contain Chocolate, and these 600 transactions include a 200 that includes Biscuits and chocolates. Using this data, we will find out the support, confidence, and lift.

1.Support

Support refers to the default popularity of any product. You find the support as a quotient of the division of the number of transactions comprising that product by the total number of transactions. Hence, we get

Support (Biscuits) = (Transactions relating biscuits) / (Total transactions)

= 400/4000 = 10 percent.

2.Confidence

Confidence refers to the possibility that the customers bought both biscuits and chocolates together. So, you need to divide the number of transactions that comprise both biscuits and chocolates by the total number of transactions to get the confidence.

Confidence = (Transactions relating both biscuits and Chocolate) / (Total transactions involving Biscuits)

= 200/400

= 50 percent.

It means that 50 percent of customers who bought biscuits bought chocolates also.

3.Lift

Consider the above example; lift refers to the increase in the ratio of the sale of chocolates when you sell biscuits. The mathematical equations of lift are given below.

Lift = (Confidence (Biscuits - chocolates)/ (Support (Biscuits)

= 50/10 = 5

It means that the probability of people buying both biscuits and chocolates together is five times more than that of purchasing the biscuits alone. If the lift value is below one, it requires that the people are unlikely to buy both the items together. Larger the value, the better is the combination.

Evaluation of Candidate Rules

In association rule mining, the evaluation of candidate rules involves assessing the quality and significance of the generated rules to determine their usefulness and relevance.
The metrics are commonly used in Candidate Rule are

1.Support: This metric measures the frequency of occurrence of a specific item in the dataset. It indicates how often a rule appears in the dataset. High support indicates that the rule is relevant to a significant portion of the dataset.

2.Confidence: Confidence measures the reliability or strength of a rule.

3.Lift: Lift measures the strength of association between the antecedent and consequent. It compares the observed support of the rule to what would be expected .

4.Leverage: Leverage computes the difference between the observed frequency of both items occurring together

5.Interest: Interest measures the interestingness of a rule by comparing the observed joint occurrence of antecedent and consequent

An Example of Candidate Rule : Transactions in a Grocery Store

Let ‘s create an example with transactions in a grocery store to generate candidate association rules.Suppose we have a set of transactions from a grocery store:

Transaction 1: {Bread, Milk, Eggs}

Transaction 2: {Bread, Diapers, Beer, Eggs}

Transaction 3: {Milk, Diapers, Beer, Coke}

Transaction 4: {Bread, Milk, Diapers, Beer}

Transaction 5: {Bread, Milk, Beer}

Let's generate candidate rules and evaluate them:

1.Frequent Itemsets:

- Let's say we find that {Bread, Beer} is a frequent itemset with support = 3/5 = 0.6.

2.Candidate Rule:

- A potential association rule could be: {Bread} => {Beer}

3.Metrics:

- **Support:** Support({Bread, Beer}) = 0.6
- **Confidence:** Confidence({Bread} => {Beer}) = Support({Bread, Beer}) / Support({Bread}) = 0.6 / Support({Bread}) = 0.6 / (4/5) = 0.75
- **Lift:** Lift({Bread} => {Beer}) = Support({Bread, Beer}) / (Support({Bread}) * Support({Beer})) = 0.6 / (4/5 * 4/5) = 1.5

4.Evaluation:

- Support of 0.6 indicates that 60% of transactions contain both Bread and Beer.
- Confidence of 0.75 means that among the transactions containing Bread, 75% also contain Beer.
- A lift of 1.5 suggests that the rule {Bread} => {Beer} has a positive association; the likelihood of buying Beer increases by 1.5 times when Bread is bought.

In this example, the rule {Bread} => {Beer} exhibits high support, confidence, and lift, indicating a strong association between purchasing Bread and purchasing Beer together in transactions.

This rule might suggest strategies like bundling Bread and Beer for promotions or placing them together in the store to potentially increase sales.

Introduction to regression and classification.

Regression

- Regression in data analytics refers to a statistical method used to model the relationship between a dependent variable (also known as the target or outcome variable) and one or more independent variables (predictors or features).
- Its primary goal is to understand how the independent variables affect the dependent variable and make predictions based on this relationship.

Components of Regression:

1. **Dependent Variable:**

   The variable that we want to predict or explain based on the independent variables. It's usually a continuous numerical value in regression analysis.

2. **Independent Variables:**

   These are the input variables used to predict or explain variations in the dependent variable. They can be numerical or categorical.

3. **Regression Models:**

   Various models are used to fit the relationship between the dependent and independent variables. Examples include linear regression, polynomial regression, ridge regression, and more.

Process of Regression Analysis:

1.Data Collection:

Gathering relevant data including the dependent and independent variables.

2.Exploratory Data Analysis (EDA):

Understanding the data, checking for correlations, distributions, and outliers.

3.Model Building:

Selecting an appropriate regression model based on the nature of the data and relationships observed in EDA.

4.Model Training:

Using historical data to estimate the parameters of the model to fit the relationship between variables.

5.Model Evaluation:

Assessing the performance of the model using evaluation metrics (e.g., MSE, RMSE, R-squared for regression) to understand how well the model fits the data.

6.Prediction and Inference:

Using the trained model to make predictions on new or unseen data and infer insights from the relationships between variables.

Classification

- Classification in data analytics refers to a supervised learning task where the goal is to categorize or label input data into predefined classes or categories.
- It involves predicting a discrete outcome or assigning data points to specific classes based on their characteristics or features.

Components of Classification:

1. **Input Data:**

   Data points with various features used to predict the class or category.

2. **Classes or Categories:**

   Discrete labels or categories that the model aims to predict or assign to input data.

3. **Classifier Algorithms:**

   Models used to learn patterns from the input data and assign class labels to new or unseen data. Examples include logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks.

Process of Classification:

1. **Data Collection and Preprocessing:**

Gathering relevant data and preparing it for analysis by cleaning, transforming, and encoding categorical variables.

2. **Feature Selection or Engineering:**

   Identifying important features that contribute to predicting the target classes.

3. **Model Selection and Training:**

   Choosing an appropriate classifier and training it using labeled data (data with known class labels).

4. **Model Evaluation:**

   Assessing the performance of the classifier using evaluation metrics such as accuracy, precision, recall, F1-score, confusion matrix, etc., on test data or through cross-validation.

5. **Prediction and Inference:**

   Using the trained model to predict the class labels of new or unseen data.

********** END OF MODULE 2 **********

MODULE III

Advanced Analytical Theory and Methods: Text Analysis - Text Analysis Steps, A Text Analysis Example, Collecting Raw Text, Representing Text, Term Frequency-Inverse Document Frequency (TFIDF), Categorizing Documents by Topics, Determining Sentiments, Gaining Insights.

Text Analysis

- Text analysis in data analysis involves the systematic examination of textual data to derive meaningful insights, patterns, or information.
- It's a process that uses various computational and analytical techniques to extract valuable knowledge from unstructured text.
- Data analysts perform text analysis to understand, organize, and summarize large volumes of textual data, which can come from sources like social media, customer reviews, surveys, emails, news articles, and more.
- Text analysis in data analysis is essential for extracting actionable insights from unstructured textual data
- It helps organizations to make informed decisions, understand customer sentiments, track trends, and gain valuable business intelligence.

Steps in Text Analysis

1.Data Collection: Gather text data from various sources such as social media, websites, documents, customer reviews, emails, or any other relevant repositories. This raw data is the foundation for analysis.

2.Data Cleaning: Preprocess the text data to ensure it's suitable for analysis. Steps in data cleaning include:

- Removing irrelevant characters, symbols, HTML tags, and special characters.
- Converting text to lowercase to ensure consistency.
- Handling missing or duplicated data.
- Correcting spelling errors if necessary.

3.Tokenization: Break the text into smaller units, such as words, phrases, or sentences, known as tokens. Tokenization helps in further analysis by breaking down the text into manageable pieces.

4.Stopword Removal: Eliminate common words that don't carry much information (e.g., "and," "the," "is") known as stopwords. Removing stopwords can improve the efficiency and accuracy of analysis.

5.Exploratory Data Analysis (EDA): Conduct initial analysis to understand the characteristics of the text data. This may involve:

- Counting word frequencies to identify common terms.
- Creating word clouds or frequency distributions.
- Analyzing the distribution of document lengths.
- Investigating patterns and anomalies within the data.

6.Sentiment Analysis: Determine the sentiment (positive, negative, neutral) of the text.

7.Topic Modeling: Discover underlying themes or topics present in the text

8.Model Building and Evaluation: If using machine learning algorithms for classification or prediction, develop models, train them on labeled data, and evaluate their performance using appropriate metrics.

9.Visualization and Interpretation: Visualize the results of analysis using charts, graphs, or other visual representations to communicate findings effectively. Interpret the insights gained from the analysis.

10.Iterative Process: Text analysis often involves an iterative approach, refining steps based on initial findings and adjusting techniques to improve accuracy and relevance.

A Text Analysis Example

Scenario: A company wants to analyze customer feedback from product reviews on an e-commerce platform to understand sentiment, identify common issues, and improve their products.

Steps in Text Analysis:

Step1 : Data Collection: Gather product reviews from the e-commerce platform. Each review consists of text written by customers expressing their opinions and experiences with the product.

Step2: Data Cleaning: Preprocess the text data:

- Remove special characters, punctuation, and HTML tags.
- Convert text to lowercase.
- Handle missing or duplicated reviews.

Step3: Tokenization and Stopword Removal: Tokenize the reviews into individual words or phrases and remove stopwords (common words like "and," "the," etc.) that do not carry significant meaning.

Step4:Sentiment Analysis: Use a sentiment analysis algorithm to determine the sentiment (positive, negative, neutral) of each review. This process classifies reviews based on the language used to express opinions.

Step5:Exploratory Data Analysis (EDA):

- Analyze word frequencies to identify commonly occurring terms in positive and negative reviews.
- Create word clouds to visualize frequently mentioned words in positive and negative contexts.
- Investigate the length of reviews and their distribution.

Step6:Topic Modeling: Apply topic modeling techniques like Latent Dirichlet Allocation (LDA) to identify common topics or themes across the reviews. This might reveal areas of concern or positive aspects frequently mentioned by customers.

Step7:Named Entity Recognition (NER): Use NER to identify and categorize specific entities mentioned in reviews, such as product features, brand names, or customer service experiences.

Step8:Text Classification: Employ text classification algorithms to categorize reviews into specific issues or product aspects (e.g., quality, shipping, customer service) to understand which areas require improvement.

Step9:Visualization and Reporting: Visualize the results using graphs, charts, or reports to communicate findings. Highlight key insights, sentiment trends, prevalent topics, and issues identified in the reviews.

Step10:Actionable Insights: Based on the analysis, the company can take actionable steps:

- Address common issues mentioned in negative reviews by improving product quality, enhancing customer service, or refining specific features.
- Highlight positive aspects identified in positive reviews to leverage strengths and improve marketing strategies.

Collecting Raw Text

- Collecting raw text data is the foundational step in text analysis.
- It involves gathering unstructured text from various sources to be processed and analyzed for insights.
- Raw text can come from diverse places such as websites, social media platforms, surveys, documents, emails, and more.
- The goal is to compile a dataset that represents the information you want to analyze

The process is done by the following ways

Identifying Data Sources: Determine the sources that contain the text data relevant to your analysis. This could include online forums, customer reviews, news articles, social media feeds, or internal documents.

Data Extraction: Use appropriate methods or tools to extract text data from these sources. This might involve web scraping for online content, API calls to retrieve data from social media platforms, or accessing databases for relevant documents.

Format Standardization: Raw text data often comes in various formats such as HTML, PDFs, plain text, or other structured formats. Standardize the format if needed, converting files into a consistent format that can be processed uniformly.

Representing Text in text analysis

- In text analysis, representing text involves transforming unstructured text data into a format that can be understood and processed by computational models or algorithms.
- This transformation is necessary to extract meaningful insights and patterns from textual information.

Here are some common methods used to represent text:

1.Bag of Words (BoW): This method represents text as a collection of words disregarding grammar and word order. Each document is represented as a vector where each dimension corresponds to a different word, and the value in each dimension represents the frequency of that word in the document

2.Term Frequency-Inverse Document Frequency (TF-IDF): TF-IDF measures the importance of a word in a document relative to a collection of documents. It assigns weights to words based on how frequently they appear in a specific document compared to their frequency across all documents in the corpus.

3.Word Embeddings: Word embeddings represent words as dense vectors in a continuous vector space

4.N-grams: Instead of considering single words, N-grams represent sequences of 'n' adjacent words.

Term Frequency-Inverse Document Frequency (TF-IDF)

Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistic used in text analysis and natural language processing to evaluate the importance of a word within a document relative to a collection of documents.

The TF-IDF technique is used for various text-based tasks:

- Document retrieval: Determining the relevance of a document to a user query.
- Text mining: Identifying keywords or important terms within a document.
- Information retrieval: Ranking and scoring documents based on their content's relevance.

Term Frequency (TF): This measures how frequently a term appears in a document. It's calculated as the number of times a term (word) appears in a document divided by the total number of words in that document. It indicates the importance of the term within the document. The weight of a term that occurs in a document is simply proportional to the term frequency.

TF(t,D) = count of t in D / number of words in D

Inverse Document Frequency (IDF): IDF measures how important a term is across a collection of documents. It's calculated as the logarithm of the ratio of the total number of documents in

the corpus to the number of documents containing the term, plus one to avoid division by zero for terms that don't appear in the corpus.

IDF(t,D) = log(N/ DF(t))

N is the total number of document

DF(t,d) is the number of document containing term t

TF-IDF Calculation: The TF-IDF score for a term in a particular document is obtained by multiplying the TF and IDF values.

TF-IDF(*t*,*d*,*D*) =TF(t,D) x IDF(t,D)

Categorizing Documents by Topics in Text Analysis

- Categorizing documents by topics is a fundamental task in text analysis, often accomplished through techniques like topic modeling.
- Topic modeling algorithms are used to automatically discover latent topics within a collection of documents
- It enables the categorization or clustering of documents based on these identified topics.
- One commonly used method for topic modeling is **Latent Dirichlet Allocation (LDA).**
- Here's an overview of how it works:

Latent Dirichlet Allocation (LDA):

1.Document-Topic and Topic-Word Probabilities:

- LDA assumes that each document in the corpus is a mixture of various topics, and each topic is a mixture of different words.
- Initially, the model randomly assigns words in the documents to topics.

2.Iterative Learning Process:

- LDA iterates to improve the assignment of words to topics until it converges to a stable solution.
- During iterations, it adjusts the assignment of words to topics based on probabilities.

3.Probability Distributions:

- The output of LDA includes two main probability distributions:
- Document-Topic distribution: For each document, the probability of belonging to various topics.
- Topic-Word distribution: For each topic, the probability distribution of words.

4.Assigning Topics to Documents:

- Once the model has converged, each document is represented as a distribution over topics based on the probabilities generated by LDA.
- Documents are categorized or clustered based on the dominant topics identified within them.

Determining Sentiments in Text Analysis

Determining sentiment in text involves assessing the emotional tone or polarity conveyed by the text—whether it's positive, negative, or neutral. This sentiment analysis process is crucial in understanding opinions, attitudes, or emotions expressed in textual data.

Sentiment Analysis Techniques:

1.Lexicon-Based Approaches:

Lexicon-based methods use dictionaries or word lists that contain sentiment scores for words. Each word is associated with a polarity (positive, negative, neutral).

2.Machine Learning-Based Approaches:

Machine learning techniques, like classification algorithms (e.g., Support Vector Machines, Naive Bayes, Neural Networks), learn to predict sentiment based on labeled training data.

3.Hybrid Approaches:

Combine lexicon-based methods with machine learning techniques to leverage the strengths of both approaches for more accurate sentiment analysis.

Gaining insights in Text Analysis

Gaining insights from text analysis involves extracting meaningful information, patterns, and knowledge from textual data to make informed decisions or understand underlying trends. Here are steps to gain insights from text analysis:

1. Exploratory Data Analysis (EDA):

- **Text Preprocessing Overview:** Clean the text data by removing noise, formatting issues, and irrelevant characters.
- **Basic Statistics:** Calculate basic statistics such as word frequencies, document lengths, or common phrases to get an initial understanding of the data.

2. Topic Modeling:

- Identify Topics: Use techniques like Latent Dirichlet Allocation (LDA) to uncover latent topics within the text documents.
- Analyze the identified topics, their associated keywords, and prevalent themes across documents to understand major content areas.

3. Sentiment Analysis:

- **Sentiment Distribution:** Determine the sentiment distribution—positive, negative, neutral—across documents or specific categories to gauge overall sentiment trends.
- **Sentiment Patterns:** Identify patterns in sentiment based on topics, time frames, or user demographics to understand sentiments related to particular aspects or events.

4. Entity Recognition and Relationship Extraction:

- **Identify Entities:** Use Named Entity Recognition (NER) to identify and categorize entities (names, locations, organizations) within the text.
- **Relationships:** Analyze relationships between entities to understand connections or associations mentioned in the text.

5. Feature Engineering and Text Representation:

- **Feature Extraction:** Create meaningful features like word embeddings or TF-IDF matrices for modeling.
- **Visualization:** Use visualizations like word clouds, bar charts, or network graphs to represent key findings, making insights more accessible.

6. Machine Learning Models:

- **Predictive Analysis:** Apply machine learning models for classification, clustering, or prediction tasks based on text features.
- **Interpret Model Outputs:** Analyze model outputs to understand what features or words contribute to predictions or classifications.

7. Feedback Loop and Iterative Analysis:

- **Iterate Analysis:** Refine the analysis based on initial findings, feedback, or additional data.
- **Continuous Improvement:** Keep updating and enhancing models or analysis techniques to improve accuracy and relevance.

************END OF MODULE 3************

MODULE IV

Advanced Analytics- Technology and Tools: MapReduce and Hadoop – Introduction to MapReduce and Apache Hadoop. The Hadoop Ecosystem – Pig, Hive, HBase, Mahout, NoSQL.**Advanced Analytics- Technology and Tools: In-Database Analytics: -** SQL Essentials – Joins, SetOperations. In-Database Text Analysis, **Data Privacy and Ethics: -** Privacy Landscape, Rights and Responsibility, Technologies.

Advanced Analytics- Technology and Tools

1. MapReduce

- MapReduce is a programming model and processing technique used in big data analytics .
- It handle and process vast amounts of data in parallel across distributed computing environments.
- It was popularized by Google and has become a fundamental concept in various data processing frameworks.
- MapReduce provides a framework for processing large-scale data by distributing the workload across multiple nodes in a cluster.
- It's particularly useful for tasks that can be parallelized, such as log processing, data mining, and various analytical tasks on large datasets.

The MapReduce process consists of two main phases:

Map Phase:

- In this phase, data is divided into smaller chunks and processed in parallel across multiple nodes in a cluster.
- Each node applies a "map" function to the data it receives, which transforms the input into intermediate key-value pairs.
- This phase breaks down the task into smaller sub-tasks that can be processed independently

Reduce Phase:

- Once the mapping phase is complete, the intermediate results are shuffled and sorted based on their keys.
- Then, the "reduce" function is applied to these intermediate key-value pairs. The reduce function aggregates, summarizes, or processes these values to generate the final output.

2.Apache Hadoop

- Apache Hadoop is an open-source framework used for distributed storage and processing of large volumes of data across clusters of commodity hardware.
- It's designed to handle massive amounts of data in a scalable and fault-tolerant manner.
- Hadoop provides a way to store, process, and analyze vast datasets that exceed the capabilities of traditional databases and processing systems.
- Hadoop's distributed nature, fault tolerance, and ability to handle large-scale data make it a foundational technology in the world of big data analytics.

Key components of the Hadoop ecosystem include:

1. **Hadoop Distributed File System (HDFS):** HDFS is a distributed file system that stores data across multiple machines in a Hadoop cluster. It provides high-throughput access to application data and is designed to be fault-tolerant.
2. **MapReduce:** MapReduce is a programming model for processing and generating large datasets in parallel across a Hadoop cluster. It allows for distributed computation of large-scale data sets across multiple nodes.
3. **YARN (Yet Another Resource Negotiator):** YARN is a resource management layer in Hadoop that manages and allocates resources across various applications running in the Hadoop cluster.
4. **Hadoop Common:** Hadoop Common contains libraries and utilities needed by other Hadoop modules. It provides support utilities, libraries, and necessary files for Hadoop modules.
5. **Hadoop ecosystem projects:** Over time, several other projects have emerged around the Hadoop ecosystem to enhance its capabilities for specific tasks. Projects like Apache Hive (for data warehousing), Apache Pig (for data flow scripting), Apache HBase (a NoSQL database), Apache Spark (for in-memory processing), and others complement Hadoop and offer various functionalities for different data processing needs.

The Hadoop Ecosystem

- *Hadoop Ecosystem* is a platform or a suite which provides various services to solve the big data problems.
- It includes Apache projects and various commercial tools and solutions.
- The components of hadoop ecosystems are

  1.PIG

  2.HIVE

  3.Mahout:

  4.NoSQL

Components of Hadoop Ecosystem are

1.PIG:

- Pig was basically developed by Yahoo which works on a pig Latin language, which is Query based language similar to SQL.
- It is a platform for structuring the data flow, processing and analyzing huge data sets.
- Pig does the work of executing commands and in the background, all the activities of MapReduce are taken care of.
- Pig Latin language is specially designed for this framework which runs on Pig Runtime. Just the way Java runs on the JVM.
- Pig helps to achieve ease of programming and optimization and hence is a major segment of the Hadoop Ecosystem.

2.HIVE:

With the help of SQL methodology and interface, HIVE performs reading and writing of large data sets.

- It's query language is called as HQL (Hive Query Language).
- It is highly scalable as it allows real-time processing and batch processing both.
- Also, all the SQL datatypes are supported by Hive thus, making the query processing easier.

- Similar to the Query Processing frameworks, HIVE too comes with two components: *JDBC Drivers* and *HIVE Command Line*.
- JDBC, along with ODBC drivers work on establishing the data storage permissions and connection whereas HIVE Command line helps in the processing of queries.

3.Mahout:

Mahout, allows Machine Learnability to a system or application.

- Machine Learning, as the name suggests helps the system to develop itself based on some patterns, user/environmental interaction or on the basis of algorithms.
- It provides various libraries or functionalities such as collaborative filtering, clustering, and classification which are nothing but concepts of Machine learning.
- It allows invoking algorithms as per our need with the help of its own libraries.

4.NoSQL

NoSQL databases, often referred to as "Not Only SQL," are a category of databases designed to handle various types of unstructured, semi-structured, or structured data in Big Data Analysis They depart from traditional relational databases (SQL databases) by offering different data models, flexibility, and scalability to manage large volumes of data efficiently.

Types of NoSQL Databases:

NoSQL databases are categorized into different types based on their data models:

1. **Document Databases:** These store data in a semi-structured format like JSON or BSON documents (e.g., MongoDB, Couchbase).
2. **Key-Value Stores:** Simplest NoSQL model, storing data as key-value pairs (e.g., Redis, Amazon DynamoDB).
3. **Column-Family Stores:** Organize data into columns and column families, suitable for large-scale distributed storage (e.g., Apache Cassandra, HBase).
4. **Graph Databases:** Designed to manage highly interconnected data, such as social networks or network topologies (e.g., Neo4j, Amazon Neptune).

Analytics- Technology and Tools: In-Database Analytics

What is In-Database Analytics

- In-database analytics is a technology that allows data processing to be conducted within the database by building analytic logic into the database itself. Doing so eliminates the time and effort required to transform data and move it back and forth between a database and a separate analytics application.
- An in-database analytics system consists of an enterprise data warehouse (EDW) built on an analytic database platform. Such platforms provide parallel processing, partitioning, scalability and optimization features geared toward analytic functionality.
- In-database analytics allows analytical data marts to be consolidated in the enterprise data warehouse.
- Data retrieval and analysis are much faster and corporate information is more secure .
- in-database analytics streamlines the analytics process, enhances performance, reduces complexity, and enables real-time or near-real-time insights generation.
- Companies use in-database analytics for applications requiring intensive processing – for example, fraud detection, credit scoring, risk management

Common examples of in-database analytics solutions include:

1. **SQL-Based Analytics:** Utilizing SQL queries with advanced analytical functions directly within the database system.
2. **Database-specific libraries:** Some databases offer libraries or extensions for machine learning, statistical analysis, and predictive modeling.
3. **Integrated Analytics Platforms:** Specialized analytical platforms or appliances that tightly integrate analytics and database capabilities for high-performance analytics.

SQL Essentials

1 Joins

- Joins in SQL are powerful operations used to combine rows from two or more tables based on related columns between them.
- They enable the retrieval of data from multiple tables simultaneously by establishing relationships between these tables.

There are different types of joins in SQL:

1.Inner Join:

- Returns rows when there is a match in both tables based on the join condition.
- SELECT * FROM table1 INNER JOIN table2 ON table1.column = table2.column;

2. Outer Join:

- Returns all rows when there is a match in either the left or right table. If there is no match, NULL values are included for columns from the opposite table.
- SELECT * FROM table1 FULL JOIN table2 ON table1.column = table2.column;

3. Left Join:

- Returns all rows from the left table and matching rows from the right table. If there is no match, NULL values are included for columns from the right table.
- SELECT * FROM table1 LEFT JOIN table2 ON table1.column = table2.column;

4.Right (Outer) Join:

- Returns all rows from the right table and matching rows from the left table. If there is no match, NULL values are included for columns from the left table.
- SELECT * FROM table1 RIGHT JOIN table2 ON table1.column = table2.column;

5.Self Join:

- When a table is joined with itself, typically used when the table contains hierarchical data or references to itself.
- SELECT e1.name, e2.name FROM employees e1 INNER JOIN employees e2 ON e1.manager_id = e2.employee_id;

2.Set Operations

- Set operations in databases are used to perform operations like union, intersection, and difference on the result sets of SQL queries.
- These operations allow data professionals to combine and manipulate data in various ways.

These set operations are handy for various scenarios in data analytics:

1.Data Integration: When combining data from multiple sources, UNION and UNION ALL help merge datasets with or without duplicates.

2.Data Validation: INTERSECT can be used to check for overlapping records between different datasets, ensuring data consistency.

3.Data Cleansing: EXCEPT or MINUS can identify data discrepancies or missing records between two datasets.

4.Data Manipulation: Set operations enable data professionals to filter and manipulate datasets in complex ways based on set theory principles.

Primary set operations:

1.UNION:

Combines the result sets of two or more SELECT statements into a single result set. It removes duplicates by default.

```
SELECT column1 FROM table1

UNION

SELECT column1 FROM table2;
```

2.INTERSECT:

Returns rows that appear in both result sets of two SELECT statements.

```
SELECT column1 FROM table1

INTERSECT

SELECT column1 FROM table2;
```

3.EXCEPT or MINUS:

Returns distinct rows from the first SELECT statement that are not present in the second SELECT statement.

```
SELECT column1 FROM table1

EXCEPT

SELECT column1 FROM table2;
```

In-Database Text Analysis

- In-database text analysis refers to performing text processing, search, and analysis directly within a database system.
- It involves using the database's capabilities to handle and analyze textual data, enabling various text-related operations without needing to extract data to external tools or platforms.
- This approach is particularly beneficial for managing and analyzing large volumes of textual data efficiently.

Here are key components and techniques involved in in-database text analysis:

1. **Full-Text Search:** Many database systems offer built-in full-text search capabilities. These functionalities allow users to perform keyword-based searches, find specific phrases or words within text fields
2. **Text Indexing:** Databases can create indexes specifically optimized for textual data, enabling faster search and retrieval operations on large volumes of text.
3. **Text Processing Functions:** Database systems may provide functions or extensions for text processing tasks, such as tokenization (splitting text into tokens/words), normalization (converting text to a standard form),
4. **Text Mining and Analytics:** In-database text analysis can include mining insights from text data, such as identifying trends, patterns, or associations within textual information.

Data Privacy and Ethics

- Data privacy and ethics are fundamental aspects of handling, managing, and utilizing data responsibly.

Privacy Landscape

1. **Protection of Personal Information:**

   Data privacy refers to the protection of sensitive and personally identifiable information (PII) of individuals. This includes names, addresses, social security numbers, health records, financial information, etc.

2. **Legal Compliance:**

   Adherence to data privacy laws and regulations, such as the GDPR (General Data Protection Regulation) in the European Union or CCPA (California Consumer Privacy Act) in California, which outline rules for collecting, storing, processing, and sharing personal data.

3. **Consent and Transparency:**

   Ensuring individuals are informed about how their data is collected, used, and shared. Obtaining explicit consent before collecting and processing their data.

4. **Data Security Measures:**

   Implementing robust security measures like encryption, access controls, data anonymization, and regular security audits to protect against unauthorized access, breaches, or data leaks.

Rights and Responsibilities

Rights and responsibilities in data privacy and ethics are crucial aspects that both individuals and organizations need to understand and uphold.

Rights:

1. **Right to Privacy:** Individuals have the right to control their personal data, including how it's collected, used, stored, and shared.
2. **Right to Access:** Individuals have the right to access their own data that's held by organizations and understand how it's being used.
3. **Right to Correction:** Individuals can request corrections or updates to inaccurate or outdated personal data.
4. **Right to Erasure (Right to be Forgotten):** Individuals can request the deletion or removal of their personal data under certain circumstances, especially if it's no longer necessary or if consent is withdrawn.
5. **Right to Data Portability:** Individuals have the right to obtain and reuse their personal data for their purposes across different services.
6. **Right to Consent:** Individuals have the right to give informed consent for the collection and processing of their data. Organizations must obtain clear and explicit consent for data usage.

Responsibilities:

1. **Data Protection and Security:** Organizations have a responsibility to implement robust data protection measures, ensuring the confidentiality, integrity, and security of individuals' data against unauthorized access, breaches, or misuse.
2. **Compliance with Regulations:** Organizations must comply with data protection laws and regulations applicable to their operations, including GDPR, CCPA, and other regional or industry-specific regulations.

3. **Transparency and Accountability:** Organizations should be transparent about their data practices, informing individuals about how their data is used and handled. They must also be accountable for their data handling practices.
4. **Data Minimization:** Collect and retain only necessary and relevant data. Avoid excessive collection or storage of personal information that isn't essential for business purposes.
5. **Ethical Use of Data:** Use data ethically and responsibly, avoiding discriminatory practices, biases, or unethical exploitation of personal information.
6. **Respecting Individuals' Rights:** Respect individuals' rights regarding their data, including providing access, facilitating corrections, honoring deletion requests, and ensuring data portability.

Emerging Technologies in Data Privacy

1. **Homomorphic Encryption:** Allowing computations to be performed on encrypted data without decrypting it, preserving data privacy during computations.
2. **Zero Trust Architecture:** Operating on the principle of "never trust, always verify," where access controls are continuously evaluated based on various factors like device health, user behavior, etc.
3. **Privacy-Preserving Technologies:** Differential privacy, federated learning, and secure multi-party computation, enabling data analysis while preserving individual privacy.

************END OF MODULE 4************