**Big Data-I**

1. **How mahout plays an important role in data mining (Hadoop)?**

**Ans: Mahout is a machine learning library that plays a crucial role in data mining within the Hadoop ecosystem. It provides scalable implementations of various machine learning algorithms, making it easier to analyze and extract valuable insights from large datasets.**

1. **Differentiate b/w Hive and Pig.**

**Ans: Hive and Pig are both high-level data processing languages in the Hadoop ecosystem, but they serve different purposes:**

**Hive: Hive is a data warehousing and SQL-like query language that allows users to write SQL-like queries (HQL) for querying and analyzing data stored in Hadoop's HDFS. It's best suited for users familiar with SQL.**

**Pig: Pig is a scripting language designed for data ETL (Extract, Transform, Load) tasks. It offers a more flexible approach to data processing, making it suitable for complex data transformations.**

1. **Whate are the Characteristics of Big Data.**

**Ans: Volume: Refers to the massive amount of data generated, collected, and stored, often beyond the capacity of traditional databases.**

**Velocity: Describes the high speed at which data is generated and the need to process it in real-time or near-real-time.**

**Variety: Indicates the diverse types of data, including structured, semi-structured, and unstructured data, such as text, images, videos, and more.**

1. **Define YARN.**

**Ans: YARN is a resource management and job scheduling component in the Hadoop ecosystem. It serves as the resource manager for Hadoop clusters, enabling efficient and centralized resource allocation and management. YARN allows multiple applications to share and utilize cluster resources more effectively, improving overall cluster utilization and performance.**

1. **Differentiate b/w HDFS and RDBS.**

**Ans: HDFS (Hadoop Distributed File System):**

* **HDFS is a distributed file system designed for storing and processing large volumes of data across clusters of commodity hardware.**
* **It is optimized for handling large files and is fault-tolerant, meaning it can handle hardware failures without data loss.**
* **HDFS is primarily used in the context of big data processing frameworks like Hadoop for distributed storage.**

**RDBMS (Relational Database Management System):**

* **RDBMS is a database management system that stores and manages data in structured tables with rows and columns.**
* **It is designed for transactional data and provides features like data integrity, ACID (Atomicity, Consistency, Isolation, Durability) properties, and SQL support.**
* **RDBMS is suitable for structured data and is commonly used in traditional enterprise applications.**

1. **Hive Meta Store.**

**Ans:**

* **The Hive Meta Store, often referred to as the Hive Meta store, is a centralized metadata repository used by Apache Hive, a data warehousing and SQL-like query language for Hadoop.**
* **It stores metadata about tables, partitions, columns, and other objects created and managed by Hive.**
* **The Hive Meta store helps in maintaining a schema for data stored in Hadoop and allows users to query and analyze data using SQL-like syntax through Hive.**

1. **Explain the ecosystem of Hadoop.**

**Ans: The Hadoop ecosystem is a collection of open-source software tools and frameworks designed to store, process, and analyze large volumes of data. Some key components of the Hadoop ecosystem include:**

* **HDFS (Hadoop Distributed File System): A distributed file system that stores data across multiple nodes in a Hadoop cluster.**
* **MapReduce: A programming model and processing framework for distributed data processing.**
* **YARN (Yet Another Resource Negotiator): A resource management and job scheduling framework.**
* **Hive: A data warehousing and SQL-like query language.**
* **Pig: A scripting language for data ETL (Extract, Transform, Load) tasks.**
* **HBase: A NoSQL database for real-time, random read/write access to data.**
* **Spark: A fast and general-purpose cluster computing framework.**
* **Sqoop: A tool for transferring data between Hadoop and relational databases.**
* **Oozie: A workflow scheduling and coordination system.**
* **Zookeeper: A distributed coordination service.**
* **Flume: A tool for ingesting and transporting large amounts of streaming data.**
* **Mahout: A machine learning library for scalable data mining.**

**These components work together to enable the processing of large-scale data, making Hadoop a powerful platform for big data analytics.**

1. **K-mean Algorithm.**

**Ans: K-means is an unsupervised machine learning algorithm used for data clustering and partitioning. It aims to group similar data points into clusters based on their features. Here's a high-level overview of the algorithm:**

**Initialization**

* **Begin by selecting the number of clusters, denoted as 'K,' that you want to identify in the dataset.**
* **Randomly initialize K cluster centroids in the feature space. These centroids act as the center points for the clusters.**

**Assignment**

* **Iterate through each data point in the dataset and assign it to the nearest cluster centroid. This assignment is typically based on a distance metric, often using Euclidean distance.**
* **Each data point now belongs to one of the K clusters.**

**Update:**

* **Recalculate the cluster centroids as the mean (average) of all data points assigned to each cluster.**
* **These updated centroids represent the new center points for their respective clusters.**

**Convergence:**

* **Repeat the assignment and update steps iteratively until one of the convergence criteria is met. Common criteria include when the centroids no longer change significantly or after a predetermined number of iterations.**
* **The final result is a set of K clusters, each containing data points that are closer to their respective cluster's centroid than to any other centroid.**

1. **Discuss the use cases of Big Data Analytics.**

**Ans** **Big Data analytics has a wide range of applications across various industries. Some common use cases include:**

* **Customer Analytics: Analyzing customer data to improve marketing, customer service, and retention.**
* **Fraud Detection: Identifying fraudulent activities in financial transactions.**
* **Predictive Maintenance: Predicting when equipment or machinery needs maintenance to avoid downtime.**
* **Healthcare Analytics: Analyzing patient data for better diagnosis and treatment.**
* **Recommendation Systems: Providing personalized recommendations in e-commerce and content streaming platforms.**
* **Supply Chain Optimization: Optimizing logistics and supply chain operations.**
* **Social Media Analysis: Analyzing social media data for sentiment analysis and market research.**
* **Cybersecurity: Detecting and preventing cyber threats.**
* **Environmental Monitoring: Analyzing sensor data for environmental research and conservation efforts.**

1. **Discuss Hive Architecture.**

**Ans: Hive is a data warehousing and SQL-like query language for Hadoop. Its architecture consists of the following components:**

* **Metastore: Stores metadata about tables, partitions, and columns.**
* **Driver: Manages query compilation, optimization, and execution.**
* **Compiler: Translates HQL (Hive Query Language) into a series of MapReduce jobs.**
* **Execution Engine: Executes the MapReduce jobs generated by the compiler.**
* **Hive CLI (Command Line Interface): Provides an interactive interface for users to submit Hive queries.**
* **Thrift Server: Allows remote clients to submit Hive queries using various programming languages.**

1. **Compare the following HDFS and GPFS.**

**HDFS (Hadoop Distributed File System):**

* **Open-source distributed file system.**
* **Designed for big data storage and processing in Hadoop clusters.**
* **Provides high fault tolerance through data replication.**
* **Optimized for sequential data access patterns.**
* **Scales horizontally by adding more commodity hardware.**

**GPFS (General Parallel File System):**

* **Proprietary parallel file system developed by IBM.**
* **Used in high-performance computing (HPC) and enterprise environments.**
* **Supports both parallel and distributed data access.**
* **Offers features like advanced data management, data compression, and snapshots.**
* **Scales vertically by adding more resources to a single server.**

1. **Explain detail about clusters and its type.**

**Ans: Clusters in the context of Big Data refer to groups or collections of data points that share common characteristics or exhibit similarities within a massive and often complex dataset. These clusters are formed through data analysis techniques and are essential for understanding patterns, trends, and structures in large-scale data.**

**Properties of Clusters in Big Data:**

* **Real-time Processing: In some cases, clusters need to be identified in real-time as data streams in, requiring streaming clustering algorithms.**
* **Spatial and Temporal Characteristics: Clusters can be based on spatial proximity, temporal patterns, or a combination of both in geospatial and time-series data.**

**Types of Clusters in Big Data: Big Data encompasses various types of data, and different clustering techniques are employed based on the data's nature and characteristics:**

* **Spatial Clustering: This type of clustering groups data points that are spatially close to each other.**
* **Temporal Clustering: Temporal clustering focuses on patterns within time-related data.**
* **Textual Clustering: Textual clustering is used for Big Data involving unstructured text data. It groups documents or text snippets with similar content, enabling topic modeling, document categorization, and sentiment analysis.**
* **Graph-based Clustering: In scenarios where data is represented as a graph or network, graph-based clustering identifies groups of nodes that are densely interconnected.**
* **Streaming Clustering: Streaming clustering algorithms continuously analyze data as it arrives in real-time, identifying clusters on the fly.**

**Applications in Big Data:**

**Clusters and their types in Big Data have numerous applications across various industries and domains:**

* **Customer Segmentation: Clustering is used to group customers based on their behavior and preferences, aiding in targeted marketing and personalized recommendations.**
* **Recommendation Systems: Clustering helps in creating recommendation engines by identifying users or items with similar characteristics and suggesting relevant content or products.**
* **Fraud Detection: Clustering can identify unusual patterns in transaction data, helping detect fraudulent activities and anomalies.**

**DWDM-I**

1. **Write any to advantages of data mart.**
2. **Name and briefly explain three type of data preprocessing methods.**
3. **List any four challenges in Data Mining.**
4. **Write the steps for efficient processing of OLAP queries.**
5. **Differentiate among ROLAP, MOLAP, and HOLAP.**
6. **Explain any four types of data cube operations with suitable example.**
7. **Explain various similarity and dissimilarity measures with suitable examples.**

**DWDM-I-R**

1. **ROLAP, MOLAP**
2. **Apriori Algo.**
3. **Data marts.**
4. **Data Mining & Challenges.**
5. **Data Preprocessing.**
6. **Architecture of data warehousing.**
7. **Data Cube / OLAP operation.**
8. **4**
9. **Naïve Algorithm & example.**
10. **Differentiate among ROLAP, MOLAP, and HOLAP.**
11. **8**
12. **Apriori Algorithm.**

**NSC-I**

1. **Define the cryptography with the help of an example.**
2. **How do you define vulnerability?**
3. **Differentiate b/w security service and security mechanism?**
4. **Write the 3 components in the CIA models?**
5. **Differentiate b/w Euclidean and extended Euclidean algorithm?**
6. **Discuss the Fermat theorem in cryptography?**
7. **Explain conventional encryption model? Discuss each component with diagram.**

**DL-I-R**

1. **Supervised and Unsupervised.**
2. **Overfitting and Underfitting.**
3. **Discuss Bais and variance.**
4. **Stochastic gradient descent.**
5. **Computational Graph.**
6. **Gradient decent, type, elaborate, etc. and it.**
7. **FFNN, architecture, application, limitations, Numerical.**

**or**

**Forward Propagation.**

**AWSN-I**

1. **Differentiate b/w Wireless Adhoc and Wireless Sensor Network?**
2. **Define me use of Routing Protocol in Wireless Networks?**
3. **List any three application of wireless sensor networks?**
4. **Differentiate b/w proactive and reactive Routing Protocols.**
5. **Discuss the concepts, Architecture and Application of Wireless Sensor Networks in details?**
6. **List issues and goals of designing Routing Protocols for Adhoc Wireless Networks in details?**
7. **Discuss the Adhoc On-Demand Distance Vector Routing Protocol in details with example?**

**Data Whare House and Data Mining-II**

1. **Define the terms Minimum Support and Confidence in context of Association rule mining.**
2. **List two advantages of Apriorism algorithm for association rule mining.**
3. **Identify and four application areas of association rule mining.**
4. **List two advantages of FP-Growth algorithm for association rule mining.**
5. **Recall the steps for Navie algorithm for association rule mining and give one example.**
6. **Draw the itemset for dynamic item Set Counting algorithm with Minium support 25% and m = 2 for the following set transactions:**
   * 1. **Transaction ID Itema Itemb Itemc**

**T1 1 1 0**

**T2 1 0 0**

**T3 0 1 1**

**T4 0 0 0.**

1. **Find Association Rules using Apriorism Algorithm with minimum support 50% and confidence 75% for the following set of transaction and also explain the algorithm.**

**Transaction ID Items**

**T1 Bread, Cornflakes, Eggs, Jam.**

**T2 Bread, Cornflakes, Jam.**

**T3 Bread, Milk, Tea**

**T4 Bread, Jam, Milk**

**T5 Cornflakes, Jam, Milk**

**NSC-II**

1. **Define DES in cryptography?**
2. **What are the block cipher design principles.?**
3. **What is digital signature and how it created.**
4. **Define Hash functions.**
5. **Explain Diffie Hellman key exchange problem with the help of an example?**
6. **Discuss RSA algorithm, Explain with example.**
7. **Explain the different Classical Cryptographic techniques in details.**