**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.

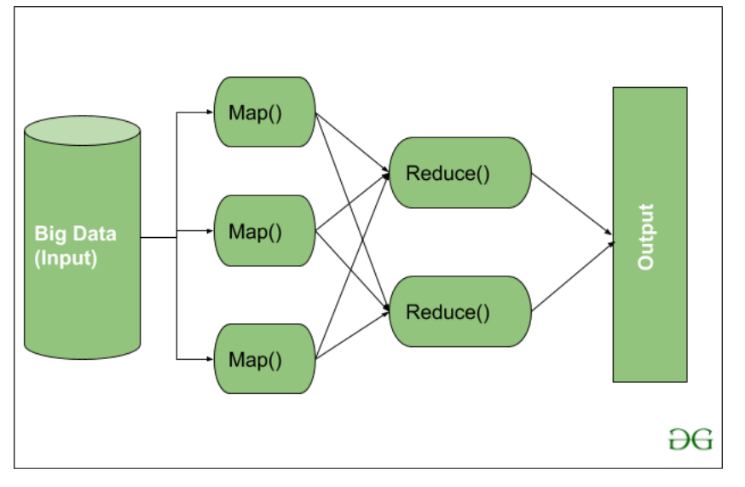
1. **Hive Meta Store.**

**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.
* Spark: A fast and general-purpose cluster computing framework.
* Sqoop: A tool for transferring data between Hadoop and relational databases.
* Zookeeper: A distributed coordination service.
* 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. **Discuss the use cases of Big Data Analytics.**

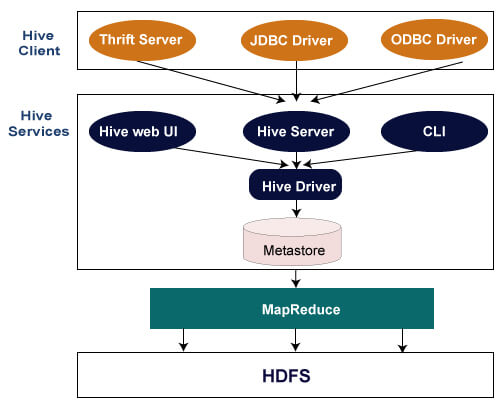
**Ans** Big Data analytics has a wide range of applications across various industries. Some common use cases include:

* 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.

**DWDM-I**

1. **Write any to advantages of data mart or Dmart.**

**Ans:** Data Marts play a vital role in data mining by providing a focused and efficient way to access and analyze data that is tailored to the specific needs of business analysts, data scientists, and other users within an organization.

Advantages of Data Mart (Dmart):

* Focused Data: Data marts contain specific, tailored data subsets, making it easier for users to access relevant information for their particular needs.
* Improved Performance: They typically offer faster query performance compared to larger data warehouses due to their smaller size and specialized focus.

1. **Name and briefly explain three type of data preprocessing methods or Data Preprocessing.**

**Ans: Data Cleaning:**

Data cleaning, also known as data cleansing or data scrubbing, is the process of identifying and rectifying errors or inconsistencies in a dataset. It involves:

* Handling missing values: Removing or imputing missing data points to avoid gaps in the dataset.
* Noise reduction: Smoothing noisy data by applying techniques like filtering or aggregation.
* Data deduplication: Removing duplicate records to ensure data integrity.

**Data Transformation:**

Data transformation involves converting the format or structure of data into a more suitable form for analysis. Key aspects of data transformation include:

* Normalization: Scaling numerical attributes to a standard range (e.g., between 0 and 1) to eliminate variations in scale.
* Encoding categorical data: Converting categorical variables into numerical representations for analysis, often through techniques like one-hot encoding.

**Data Reduction:**

Data reduction methods aim to reduce the volume of data while preserving its quality and meaningful information. These methods include:

* Sampling: Selecting a representative subset of data points from a larger dataset, which can significantly reduce computational requirements for analysis.
* Histogram analysis: Analyzing data distributions to identify patterns and reduce data size by capturing the most relevant aspects of the data.

These data preprocessing methods are crucial for preparing data for analysis in data mining, machine learning, and other data-driven tasks, as they ensure that the data is clean, well-structured, and suitable for accurate modeling and insights extraction.

1. **List any four challenges in Data Mining. Or Data Mining & Challenges.**

Ans: **Data Quality:**

Data in real-world applications often contain errors, missing values, inconsistencies, and noise. Low-quality data can lead to inaccurate or unreliable results in data mining. Ensuring data quality is a fundamental challenge in the field.

**Scalability:**

Data mining algorithms must handle increasingly large datasets, often with millions or billions of records.

**Privacy and Security:**

Data mining involves the analysis of sensitive and personal information. Ensuring the privacy and security of data while extracting valuable insights is a significant challenge.

**Complexity of Algorithms:**

Many data mining algorithms, such as neural networks, decision trees, and association rule mining, can be complex and computationally intensive. Implementing and fine-tuning these algorithms can be challenging, especially for non-experts in the field.

1. **ROLAP, MOLAP.**

ROLAP (Relational Online Analytical Processing) is a data storage and querying approach often used in data mining and business intelligence applications. In a nutshell:

**Data Storage:** ROLAP stores data for data mining in relational databases. It utilizes the tabular structure of relational databases to organize and manage data. This allows data miners to work with structured and well-organized data that can be easily queried using SQL and other relational database tools.

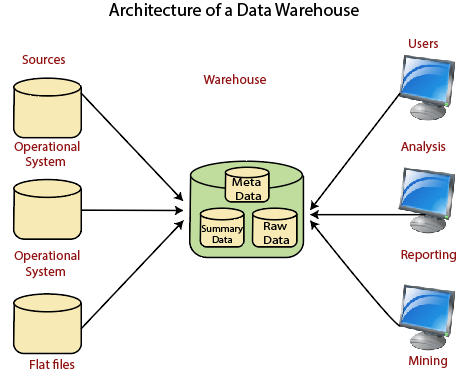
**Querying:** Data mining tasks often require complex querying and analysis of large datasets.

**MOLAP (Multidimensional Online Analytical Processing)**: is a data storage and querying approach frequently employed in data mining and business intelligence applications. Here are its key aspects in data mining:

**Data Storage (1 Mark):** MOLAP stores data for data mining in specialized multidimensional cubes or structures. These cubes are designed to efficiently represent and store multidimensional data, making them highly suitable for complex data mining tasks. The data in MOLAP systems is pre-aggregated and optimized for rapid retrieval and analysis.

**Querying (1 Mark):** Data mining often involves querying and analyzing large volumes of data from various dimensions and measures.

1. **Apriori Algo.**

**Ans:** Apriori algorithm refers to the algorithm which is used to calculate the association rules between objects. It means how two or more objects are related to one another. In other words, we can say that the apriori algorithm is an association rule leaning that analyzes that people who bought product A also bought product B.

1. **Architecture of data warehousing.**

Ans: A data warehouse architecture is a method of defining the overall architecture of data communication processing and presentation that exist for end-clients computing within the enterprise. Each data warehouse is different, but all are characterized by standard vital components.

1. **Data Cube / OLAP operation**

**|Ans: Data Cube (1 Mark):** A data cube is a multidimensional structure that stores data in a way that allows for efficient and flexible analysis. It typically includes dimensions (attributes), measures (data values), and hierarchies. Data cubes are essential for performing multidimensional analysis in OLAP systems.

**OLAP Operations (1 Mark):** OLAP operations are actions applied to data cubes for analysis. Common OLAP operations include roll-up (aggregating data to higher levels), drill-down (exploring more detailed data), pivot (changing the viewpoint), slice (selecting a specific subset), dice (selecting a subcube), drill-through (accessing detailed data), ranking (assigning ranks to data elements), and filtering (applying conditions). These operations provide users with the flexibility to explore and derive insights from multidimensional data efficiently.

1. **Differentiate among ROLAP, MOLAP, and HOLAP.**

Ans: **ROLAP (Relational Online Analytical Processing):** ROLAP is an OLAP approach that stores data in relational databases, using tables and SQL queries to represent and analyze multidimensional data. It provides flexibility for complex data models and diverse data sources.

**MOLAP (Multidimensional Online Analytical Processing):** MOLAP is an OLAP approach that stores data in specialized multidimensional cubes, pre-aggregated and optimized for efficient multidimensional data analysis. It excels in query performance and interactive analysis.

**HOLAP (Hybrid Online Analytical Processing):** HOLAP combines elements of ROLAP and MOLAP by allowing users to store summary data in multidimensional cubes for performance while retaining detailed data in relational databases for flexibility. It offers a balance between flexibility and query speed.

1. **Explain any four types of data cube operations with suitable example.**

**Ans:** Cube operations refer to various actions and manipulations performed on data cubes in the context of Online Analytical Processing (OLAP) to analyze and extract meaningful insights from multidimensional data. Here are some common cube operations:

**Roll-Up (Drill-Up):** Roll-up means looking at data from a higher, summary level. It's like going from detailed monthly sales to quarterly or yearly sales.

**Drill-Down (Roll-Down):** Drill-down is the opposite; it's diving into more detailed data. It's like going from yearly sales to monthly or daily sales.

**Slice:** Slicing is like cutting out a specific part of data. It's selecting data for a particular category or time period, such as "sales for a specific product in a given month."

**Dice:** Dicing is like taking a smaller piece of a pizza. It's selecting data for a combination of categories or conditions, such as "sales for a specific product in a given month and region."

1. **Naive Algorithm & example**
2. **Apriori Algorithm.**

**NSC-I**

1. **Define the cryptography with the help of an example. 4-2**

**Ans:** Cryptography is the practice and study of techniques used to secure communication and information from unauthorized access or modification. It involves the use of mathematical algorithms and principles to transform plaintext (unencrypted data) into ciphertext (encrypted data) in order to protect its confidentiality, integrity, and authenticity.

Exp: Original Message (Plaintext): "HELLO"

Shift Value: 3

Encrypted Message (Ciphertext): "KHOOR"

Encrypted Message (Ciphertext): "KHOOR"

Shift Value: 3 (opposite of the encryption shift)

Decrypted Message (Plaintext): "HELLO".

1. **Define DEC in cryptography. -2**

**Ans: DEC** stands for "Data Encryption Standard." It is a symmetric-key block cipher that was developed by the United States National Bureau of Standards (now known as the American National Standards Institute) in the 1970s. The DEC algorithm is widely used for encrypting data in various applications, including secure communication protocols like SSL/TLS and PGP.

1. **What are the block chipper design principles. -2**

Block Cipher is an encryption algorithm that works with a symmetric key in a deterministic way. The plain text is divided into several blocks of equal size. If the length of the plain text does not allow block division of equal size, padding is done over the plain text. His type of encryption method can encrypt on blocks of 128 bits, the key can be 128, 192, or 256 bits.

**Principals:**

**The number of encryption rounds** − the number of encryption rounds that the plain text will go through explains the decoding difficulty and hence establishes security.

**Key scheduling algorithm** − the generation of the keys for each of the rounds is defined by this algorithm.

**Function Design**.

1. **How do you define vulnerability? -2**

**Ans:** A vulnerability is a weakness that can be exploited by cybercriminals to gain unauthorized access to a computer system. After exploiting a vulnerability, a[cyberattack](https://www.upguard.com/blog/cyber-attack) can run malicious code, install[malware](https://www.upguard.com/blog/malware), and even steal[sensitive data](https://www.upguard.com/blog/sensitive-data).

1. **Difference between AEC and DES. 4-2**

|  |  |  |
| --- | --- | --- |
| 1. | AES stands for [Advanced Encryption Standard](https://www.geeksforgeeks.org/advanced-encryption-standard-aes/) | DES stands for [Data Encryption Standard](https://www.geeksforgeeks.org/data-encryption-standard-des-set-1/) |
| 2. | The date of creation is 2001. | The date of creation is 1977. |
| 3. | Byte-Oriented. | Bit-Oriented. |
| 4. | Key length can be 128-bits, 192-bits, and 256-bits. | The key length is 56 bits in DES. |
| 6. | It is faster than DES. | It is slower than AES. |
| 8. | It is flexible. | It is not flexible. |

Ans:

1. **Differentiate b/w security service and security mechanism?4-2**

**Ans:** Security services are high-level functions that provide protection to computer systems and networks. These services are typically provided by a security system or a security application, and they include:

1. Authentication: The process of verifying the identity of a user, device, or system.
2. Authorization: The process of granting or denying access to resources based on a user's identity and privileges.
3. Confidentiality: The protection of sensitive information from unauthorized access, use, or disclosure.

Security mechanisms are the specific techniques and technologies used to provide security services. These mechanisms can be hardware-based, software-based, or a combination of both. Some common security mechanisms include:  
  
1. Firewalls: Network devices that control incoming and outgoing network traffic based on a set of security rules.  
2. Encryption: The process of converting plaintext data into unreadable ciphertext to protect it from unauthorized access.

1. **Write the 3 components in the CIA models?4-2**

**Ans:** The CIA model, also known as the CIA triad, is a widely accepted framework used in information security to ensure the confidentiality, integrity, and availability of data. These three components are essential for maintaining the security and protection of sensitive information.  
  
1. **Confidentiality**: Confidentiality refers to the assurance that information is only accessible to authorized individuals or entities.   
2. **Integrity**: Integrity focuses on the accuracy, consistency, and trustworthiness of data throughout its lifecycle.   
**3. Availability:** Availability ensures that information and resources are accessible and usable when needed by authorized users.

1. **Differentiate b/w Euclidean and extended Euclidean algorithm? -4**

**Ans: The Euclidean algorithm** is a method for finding the greatest common divisor (GCD) of two integers, and it is based on the principle of iterative division. The algorithm starts with two integers, a and b, and repeatedly divides the larger number (b) by the smaller number (a) until the quotient is 0. The remaining number (b) is the GCD of a and b.

**The extended Euclidean algorithm** is an extension of the Euclidean algorithm that allows us to find the solution of a system of linear equations ax + by = c, where a, b, and c are integers. The algorithm involves a series of transformations that convert the system of equations into a set of simpler equations, which can be solved using the Euclidean algorithm.

**1 GCD Computation**:

The Euclidean algorithm is limited to finding the GCD of two integers.

while the extended Euclidean algorithm can be used to find the GCD of two integers as well as to solve systems of linear equations.

2. **Number of Operations**:

The Euclidean algorithm typically requires fewer operations than the extended Euclidean algorithm, especially for larger inputs.

However, the extended Euclidean algorithm can be more efficient for certain types of inputs, such as when the coefficients of the system of linear equations are relatively prime.

3. **Applications**:

The Euclidean algorithm has a wider range of applications, including cryptography, coding theory, and computer science,

while the extended Euclidean algorithm is primarily used in number theory and algebraic geometry.

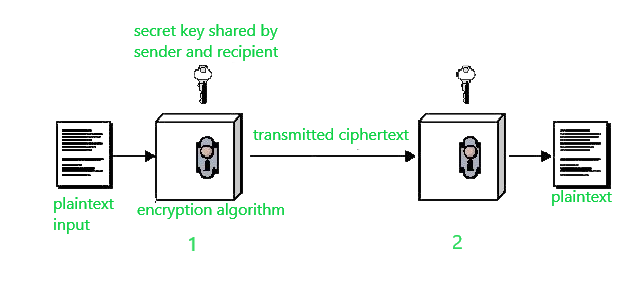
1. **Discuss the Fermat theorem in cryptography?-4**

**Ans:** In cryptography, the Fermat theorem is used in the design of cryptographic protocols, particularly in the area of public-key cryptography. Public-key cryptography relies on the difficulty of certain mathematical problems, such as factoring large numbers and computing discrete logarithms, to ensure the security of the encrypted data. The Fermat theorem is used in the design of these cryptographic protocols because it provides a way to prove the security of the system without having to explicitly compute the discrete logarithm of a number.

a^(p-1) ≡ 1 (mod p)

1. **Explain conventional encryption model? Discuss each component with diagram.8**

**Ans:** The conventional encryption model, also known as symmetric encryption or secret key encryption, is a widely used method of encrypting and decrypting data. In this model, the same key is used for both encryption and decryption processes. The key is kept secret and shared only between the sender and the receiver.

****

Conventional encryption has mainly 5 ingredients:

Plain text –   
It is the original data that is given to the algorithm as an input.

Encryption algorithm –   
This encryption algorithm performs various transformations on plain text to convert it into ciphertext. 

Secret key –   
The secret key is also an input to the algorithm. The encryption algorithm will produce different outputs based on the keys used at that time. 

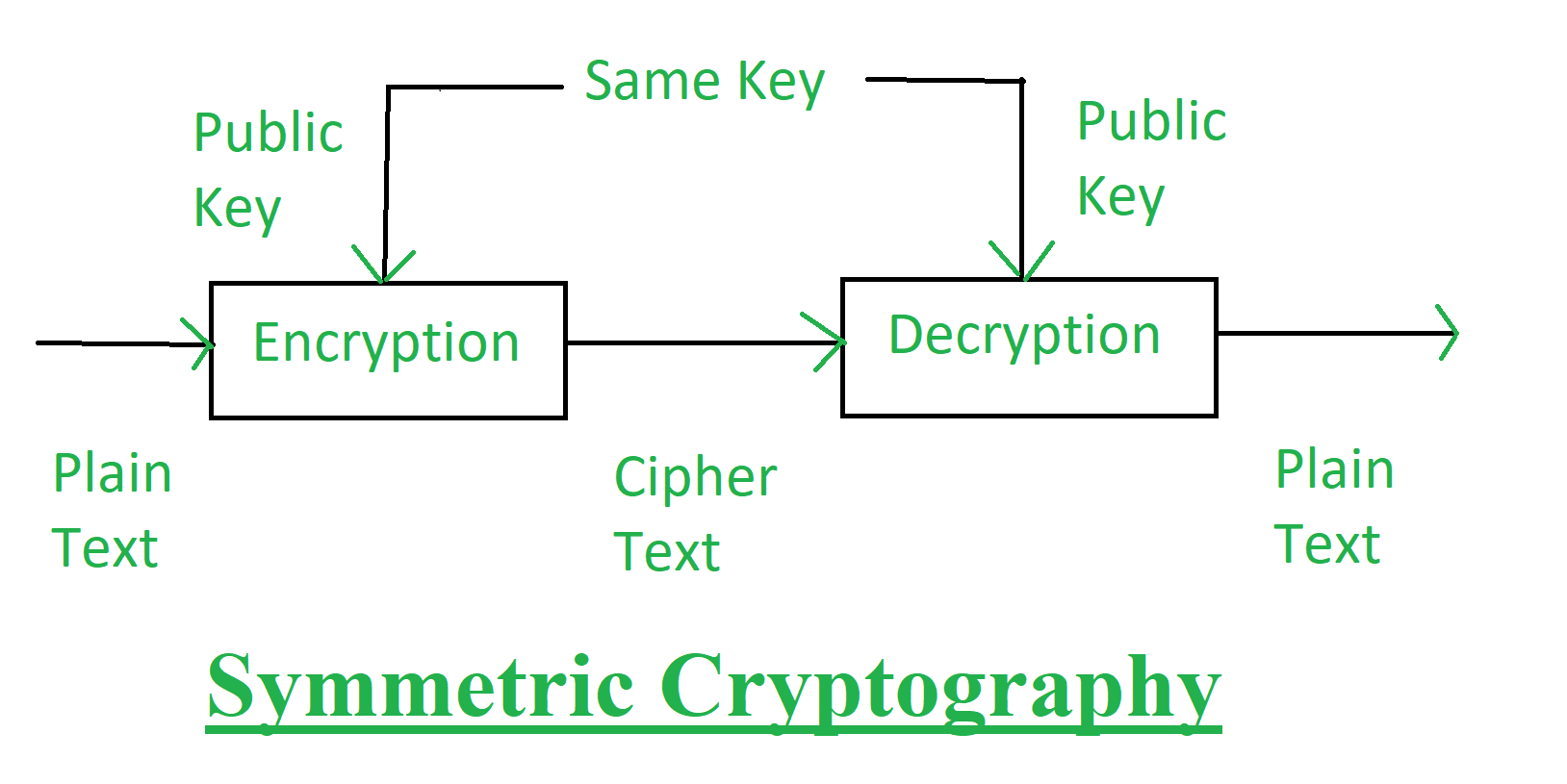
Ciphertext –   
It contains encrypted information because it contains a form of original plaintext that is unreadable by a human or computer without proper cipher to decrypt it. It is output from the algorithm. 

Decryption algorithm –   
This is used to run encryption algorithms in reverse. Ciphertext and Secret key is input here and it produces plain text as output.

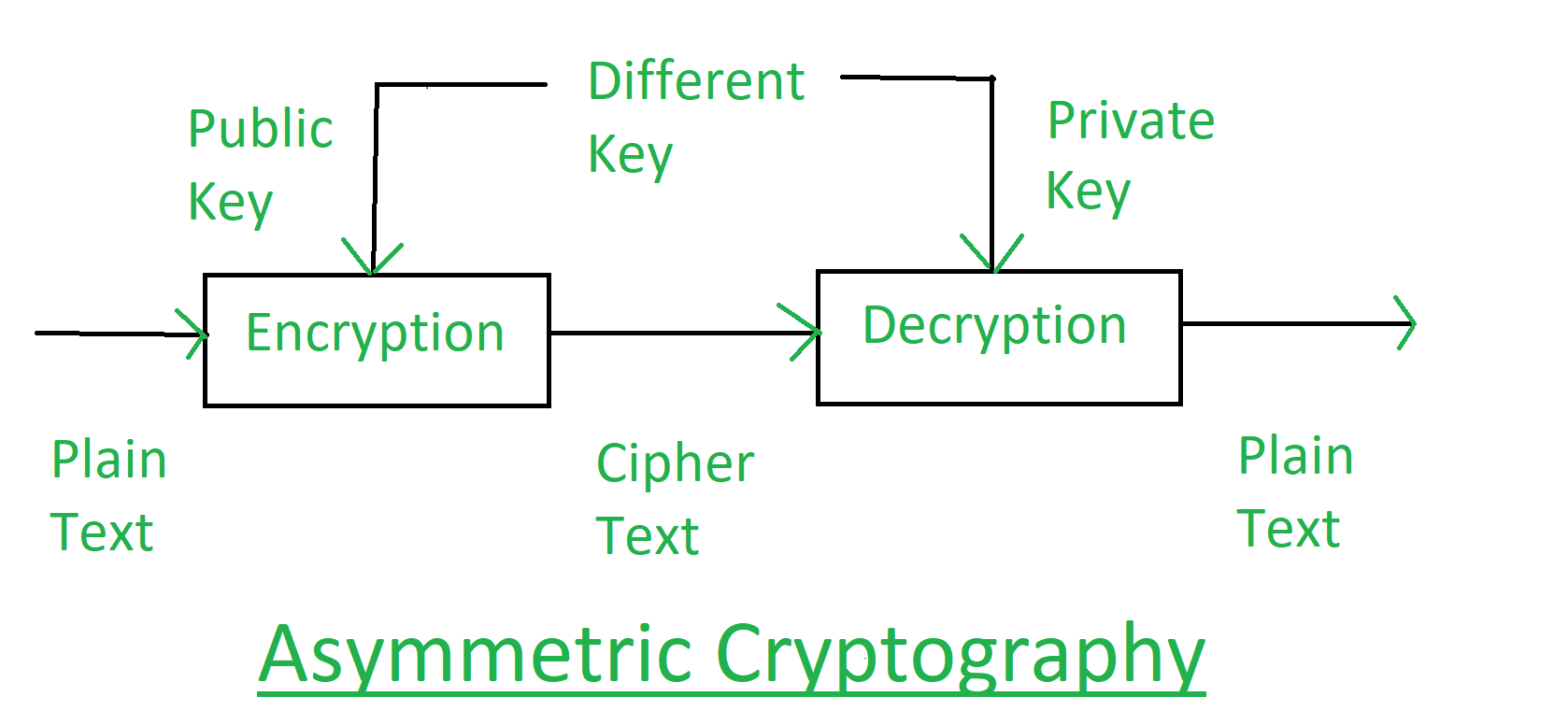
1. **Explain the different classical cryptography technique in details.8**

**Ans: Classical cryptography** refers to the traditional methods of encrypting and decrypting information that were used before the advent of computers and modern cryptographic techniques. It primarily relies on symmetric key algorithms, where the same key is used for both encryption and decryption.

**Symmetric cryptography,** also known as secret-key cryptography, is a cryptographic method where the same key is used for both encryption and decryption of data. In this approach, the sender and receiver must share the same secret key in advance.



**Asymmetric cryptography**, also known as public-key cryptography, is a cryptographic method that uses two different but mathematically related keys: a public key and a private key. These keys are generated as a pair, where what one key encrypts, only the other corresponding key can decrypt.  
In asymmetric cryptography, each user has a unique pair of keys. The public key is freely distributed to anyone who wants to communicate with that user, while the private key remains confidential and known only to its owner. The public key can be used to encrypt messages, while the private key is used for decryption.



**DL-I-R**

1. **Difference between overfitting & underfitting. - 2**

**Ans:**

|  |  |  |
| --- | --- | --- |
|  | Over Fitting | Under Fitting |
| Complexity of the Model | More complex model | More simple model. |
| Regularization | Less Regularization | More regularization |
| Quantity of features | A large quantity of features | A smaller quantity of features. |
| Data | Data cleaning, cross validation. | Data cleaning, cross validation. |

1. **Difference between forward & Back Propagation.**

**Ans:**

|  |  |  |
| --- | --- | --- |
| Aspect | Forward Propagation | Back Propagation |
| Process Definition | Computing the output of a neural network given an input. | Updating the network's parameters based on prediction error. |
| Usage Phase | Training and inference. | Primarily during training. |
| Computational Complexity | Relatively straightforward, single pass. | More complex, multiple passes and calculations. |
| Parameter Adjustment | No parameter adjustments. | Parameters adjusted to learn from errors. |
| Information Flow Type | Feed-forward, no feedback. | Introduces feedback through error propagation. |

1. **Gradient Descent.**

**Ans:** Gradient descent is an optimization algorithm used to minimize the loss function in machine learning. It is a first-order optimization algorithm that iteratively adjusts the parameters of a model to find the values that minimize the loss function.

**Steps:** 1- Initialize, 2- Compute for loss function. 3-Compute for Gradient, 4-Update.

1. **Define Bias & variance.**

**Ans: Bias** is the difference between the actual value and the predicted value.

**Variance:** The amount by which the model prediction would change if we estimate it using a different dataset.

1. **Computational graph.**

**Ans:** A computational graph, also known as a computation graph or a directed acyclic graph (DAG), is a graphical representation of mathematical operations and their dependencies in deep learning models. It is a fundamental concept used in deep learning frameworks to represent and optimize the computations involved in training and inference processes.  
  
1**. Structure:** A computational graph consists of nodes and edges. Nodes represent mathematical operations or variables, while edges represent the flow of data between these operations. Each node performs a specific operation on its inputs and produces an output.

**2. Forward and backward propagation:** Computational graphs are particularly useful for implementing forward and backward propagation algorithms, which are essential for training deep learning models.

3**. Efficiency and parallelism:** Computational graphs enable efficient execution of deep learning models by allowing parallelization of computations.

4. **Automatic differentiation:** Another key advantage of computational graphs is their ability to automatically compute gradients using automatic differentiation techniques. By representing mathematical operations as nodes in the graph.

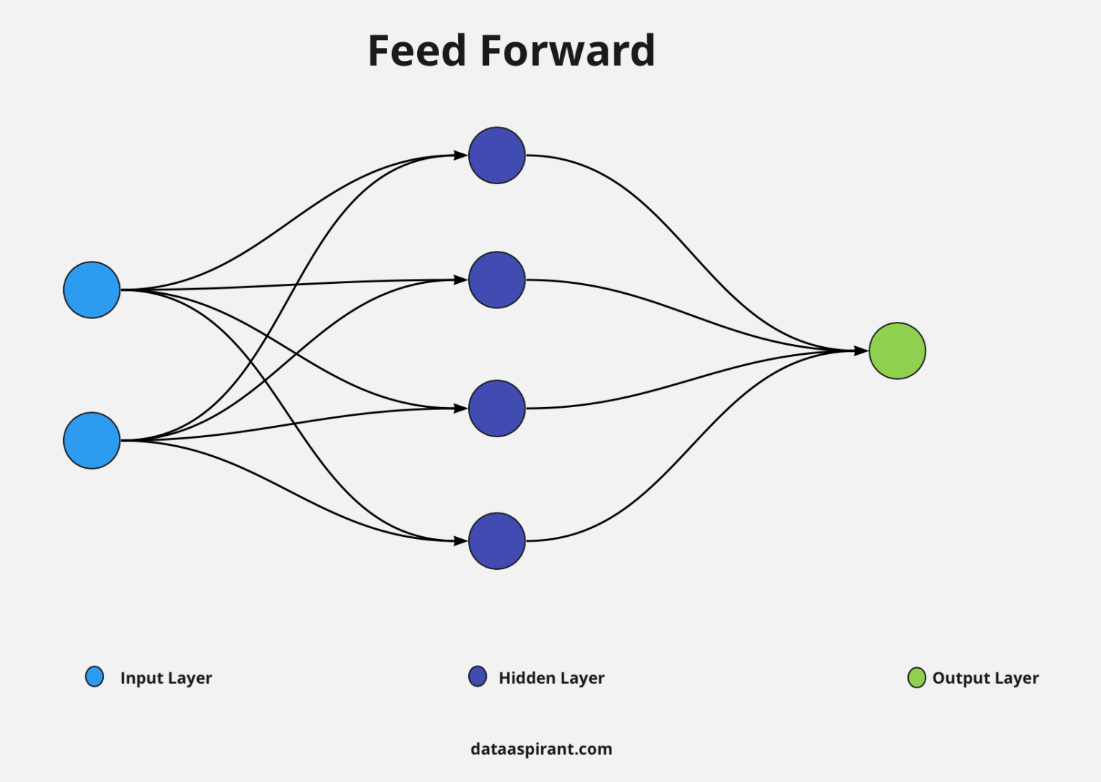
1. **Difference between supervised & unsupervised learning.**

**Ans:**

|  |  |  |
| --- | --- | --- |
|  | Supervised Learning | Unsupervised Learning. |
| Data Labeling | 1. Required labeled data | Work with unlabeled data. |
| Object | 1. Predicting outputs based on inputs. | Discovering patterns and structures. |
| Training Process | 1. Minimizing prediction errors with labels. | Find Patterns or grouping data. |
| Aspect | 1. Supervised Learning | Unsupervised. |
| Data Preparation | 1. Requires Labeling | Working with raw, unlabeled data. |
| Guided Learning | 1. Guided by explicit labels | Operates without explicit guidance |

1. **Feed forward neural networks.**

**Ans:** Feed forward neural networks are a type of artificial neural network where the information flows in one direction, from input layer to output layer, without any feedback loops. This type of network is widely used in various applications such as image recognition, speech recognition, and natural language processing.



**Input Layer:** It is starting layer of the network that has a weight associated with the signals.

**Hidden Layer:** This layer lies after the input layer and contains multiple neurons that perform all computations and pass the result to the output unit.

**Output Layer:** It is a layer that contains output units or neurons and receives processed data from the hidden layer, if there are further hidden layers connected to it then it passes the weighted unit to the connected hidden layer for further processing to get the desired result.

|  |  |
| --- | --- |
| **Advantages** | **Dis-advantages** |
| Simple to Implement | Limited Memory |
| Fast Training | Difficult to Interpret |
| Flexible | Overfitting |

**Application:**

* Natural Language Processing
* Speech Recognition
* Image Recognition

**Limitations:** It has some limitations like sometimes information about the neighborhood is lost and, in that case, it becomes difficult to process further all steps are needed to be performed again and it does not support back propagation so the network cannot learn or correct the fault of the previous stage.

**AWSN-I**

1. **Write AWSN introduction and its application.**
2. **Differentiate or difference b/w Wireless Adhoc and Wireless Sensor Network?**
3. **Write difference between WSN and Wired Networks.**
4. **Define DSDV and example with table.**
5. **MAC protocols and its types (LEACH, S-MAC).**
6. **Discuss the concepts, Architecture and Application of Adhoc Wireless Sensor Networks in details?**
7. **Define Routing Protocol and its type (Two Categories) in Adhoc Wireless Networks with example.**
8. **Define me use of Routing Protocol in Wireless Networks?**
9. **List any three application of wireless sensor networks?**
10. **Differentiate b/w proactive and reactive Routing Protocols.**
11. **Discuss the concepts, Architecture and Application of Adhoc Wireless Sensor Networks in details?**
12. **List issues and goals of designing Routing Protocols for Adhoc Wireless Networks in details?**
13. **Discuss the Adhoc On-Demand Distance Vector Routing Protocol in details with example?**
14. **AWSN introduction & application. ----------------------**
15. **WSN and W Adhoc Networks difference. ----------------**
16. **WSN and wired Networks difference------------------------**
17. **DSDV example with Table. ------------------------**
18. **MAC protocols (LEACH, S-MAC).-----------------------**
19. **Architecture of AWSN.-----------------------------------**
20. **Routing Protocols (two categories)------------------------**

**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.**

**05-11-2023**

**BD-MST-II**

1. **How classification algorithm is different from linear regression algorithm. -2**

**Ans:** Classification algorithms and linear regression algorithms are two different types of machine learning techniques used for different types of data. Classification algorithms are used when the output variable is a category, and the goal is to categorize new observations based on the patterns learned from the training data. Linear regression algorithms, on the other hand, are used when the output variable is continuous, aiming to establish a relationship between the input variables and the continuous output, thus predicting new values based on this relationship.

1. **State the concept of the segmentations. 👍**

**Ans:** Segmentation, in the context of data analysis and marketing, refers to the process of dividing a larger market or dataset into smaller, distinct groups or segments based on specific characteristics or criteria. The goal of segmentation is to identify homogeneous subgroups within a larger population, allowing for more targeted and effective marketing, analysis, or decision-making. Segmentation is a common practice in various fields, including marketing, customer relationship management, and data analysis.

1. **What is use of JAQL in ml.**

**Ans:** To create an inverted index using JAQL, you would typically use JAQL's JSON query language, which is designed for processing semi-structured and structured data. Here is an example of how you can create a simple inverted index using JAQL:

Suppose you have a JSON dataset consisting of documents with text content. Each document has an ID and some text. You want to create an inverted index that maps each word to the list of document IDs where it occurs

1. **Difference between lazy linear and eager linear. 👍**

Ans:

|  |  |  |
| --- | --- | --- |
| **Property** | **Lazy Learning** | **Eager Learning** |
| **Training Speed** | Fast, stores the data while training | Slow, Tries to learn from data while training |
| **Prediction Speed** | Too Slow tries to apply functions and learnings in the prediction stage | Faster, predicts very fast as there are pre-defined functions |
| **Learning Scope** | Medium, it can learn from data while training | Medium, it can learn from data while testing |
| **Pre-Calculated Algorithm** | Absent, calculations are done while the testing phase | At present, here calculations are already done in the training phase |

1. **Discuss real time sentiment analysis.**

**Ans:** Real-time sentiment analysis is a process that involves the use of natural language processing (NLP) and text analysis techniques to determine the emotional tone and sentiment expressed in textual data in real time. This technique is particularly useful for monitoring and understanding public opinion, customer feedback, and social media conversations as they unfold.

1. **Define Thread pooling. 👍**

**Ans:** Thread pooling is a programming technique that involves creating and managing a pool of threads to execute multiple tasks efficiently. By reusing threads from the pool, the overhead of creating and destroying threads for each task is reduced, leading to improved performance and resource management in concurrent applications.

1. **Describe the challenges for handler streaming data from real world.**

**Ans:** Handling streaming data from the real world poses various challenges, including managing high data volumes, ensuring real-time processing, maintaining data accuracy, and minimizing latency for timely analysis and decision-making. Additionally, ensuring data security and integrity, dealing with data quality issues, and managing the scalability of the system are also significant challenges in handling streaming data from the real world.

1. **Discuss decision tree. 👍**

**Ans:** Decision trees are predictive models used in machine learning for both classification and regression tasks. They work by recursively splitting the data based on input features to make decisions, creating a tree-like graph structure. Decision trees are popular due to their interpretability, simplicity, and ability to handle both numerical and categorical data. They are widely used in various fields, including finance, healthcare, and marketing.

1. **Explain the benefits of stream processing in big data world. -4 👍**

**Ans:** Stream processing is a powerful technology that enables real-time processing of high-volume, high-velocity, and high-variety data streams. In the context of big data, stream processing offers numerous benefits that help organizations make better decisions, improve operational efficiency, and gain a competitive edge. Here are some of the key benefits of stream processing in the big data world:  
  
1. **Real-time insights**: With stream processing, organizations can gain real-time insights into their data streams, enabling them to react quickly to changing conditions, detect anomalies, and make timely decisions.  
2. **Low latency**: Stream processing is designed to handle high-volume, high-velocity data streams in real-time, which means that it can process data much faster than traditional batch processing methods.

3. **Improved decision-making**: Stream processing enables organizations to make better decisions by providing real-time insights into their data streams. This can lead to improved operational efficiency, reduced costs, and increased revenue.

4. **Cost savings**: Stream processing can help organizations reduce costs by minimizing the need for data storage and processing.

| Aspect | Stream Processing | Traditional Processing |
| --- | --- | --- |
| Data Handling | Handles continuous data streams in real-time without storage | Processes static datasets stored in databases or warehouses |
| Processing Speed | Operates at high speeds for real-time data analysis and immediate responses | May have slower processing speeds, especially for large data volumes |
| Use Cases | Suitable for real-time analytics, fraud detection, IoT data analysis | Commonly used for batch processing, data warehousing, and historical data analysis |
| Scalability | Designed for horizontal scalability to handle large data volumes and increased workloads | May have limitations in scalability and may require significant hardware upgrades for increased workloads |

1. **Explain the use cases for k-mean Algorithm. 👍**

**Ans:** The k-means algorithm is a popular unsupervised machine learning algorithm used for clustering analysis. It aims to partition a given dataset into k clusters, where each data point belongs to the cluster with the nearest mean value. The algorithm iteratively assigns data points to clusters and updates the cluster centres until convergence is achieved. K-means has various use cases across different domains due to its simplicity and efficiency.  
  
1. **Image Compression:**  
One of the primary applications of k-means is image compression. Images often contain a large number of colors, but not all colors are equally important for human perception. By applying k-means clustering on the color space of an image, we can reduce the number of colors used while preserving visual quality. The algorithm groups similar colors together and replaces them with the mean color value of each cluster. This reduces the storage space required for representing images without significant loss in visual fidelity.  
  
2. Customer Segmentation:  
K-means clustering is widely used in marketing and customer analytics for segmenting customers based on their behaviour, preferences, or demographics. By analyzing customer data such as purchase history, browsing patterns, or demographic information, businesses can identify distinct groups of customers with similar characteristics. These segments can then be targeted with personalized marketing campaigns or product recommendations, leading to improved customer satisfaction and higher conversion rates.  
  
3. Anomaly Detection:  
K-means can also be utilized for anomaly detection in various domains such as network security, fraud detection, or manufacturing quality control. By clustering normal instances of a system or process, any data point that significantly deviates from its assigned cluster can be considered an anomaly or outlier. This approach helps in identifying unusual patterns or behaviours that may indicate potential security breaches, fraudulent activities, or manufacturing defects.  
  
5. Recommendation Systems:  
K-means clustering can be used in recommendation systems to group users or items based on their preferences or attributes. By clustering users with similar tastes, personalized recommendations can be generated for individuals within the same cluster. Similarly, clustering items with similar features allows for recommending similar products or content to users based on their preferences.

1. **Explain in details about clustering and its various types. -8 👍**

**Ans:** Clustering is a fundamental technique in unsupervised machine learning that involves grouping a set of data points or objects based on their inherent characteristics or similarities. The goal of clustering is to identify patterns, similarities, or associations within a dataset without any predefined labels or target values. This process helps in understanding the natural grouping or structure of the data, enabling insights into the underlying relationships and distributions within the dataset.

Clustering algorithms partition the data into subsets, known as clusters, with each cluster ideally containing data points that are more similar to each other compared to data points in other clusters. The clusters can be formed based on various metrics, such as distance, density, or connectivity, depending on the specific algorithm and the nature of the data.

Some popular clustering algorithms include:

* **K-means Clustering:** A partitioning method that divides data points into K clusters, where each data point belongs to the cluster with the nearest mean.
* **Hierarchical Clustering**: An approach that creates a hierarchy of clusters, either by recursively merging or splitting clusters based on the distance between data points.
* **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** A density-based algorithm that identifies clusters as areas of high density separated by areas of low density.
* **Mean Shift Clustering:** A method that identifies cluster centres as areas of high density and iteratively shifts points towards the mode of the data distribution.

**Gaussian Mixture Models (GMM):** A probabilistic model that assumes the data points are generated from a mixture of several Gaussian distributions, allowing for the identification of clusters based on the distribution of the data

1. **Discuss the concept of Indexing and its techniques. 👍**

**Ans:** In the context of databases and information retrieval, indexing is the process of organizing and structuring data to enhance the efficiency of data retrieval operations. It involves creating a reference or pointer to the data, making it easier to locate and retrieve specific information from a large dataset. Indexing is widely used in various applications, including search engines, databases, and information management systems.

Indexing Techniques:

* **B-Tree Indexing**: B-tree indexing is commonly used in database systems to organize and store data in a hierarchical tree structure, allowing for efficient retrieval, insertion, and deletion operations.
* **Hashing**: Hashing involves mapping data to a fixed-size array, enabling quick data retrieval based on the calculated hash values. It is often used for rapid data lookup operations.
* **Inverted Indexing**: Inverted indexing is a technique used in text retrieval systems, where a data structure is built to map keywords or terms to the documents or web pages containing them. It is widely used in search engines to facilitate fast keyword-based searches.
* **Binary Search Indexing**: Binary search indexing is used for sorted datasets, allowing for efficient retrieval by repeatedly dividing the search interval in half until the desired data is found.
* **Suffix Tree Indexing**: Suffix tree indexing is used for efficient string searching and pattern matching. It involves constructing a tree-based data structure that stores all the suffixes of a given text string.
* **Bitmap Indexing**: Bitmap indexing is used for efficient querying and analysis of large datasets by representing data in the form of bitmaps, where each bit corresponds to a specific value or attribute.

1. **Describe RTAP and some widely used RTAP. 👍**

**Ans:** TAP, which stands for Real-Time Analytics Platform, is a software system that enables organizations to process and analyze data in real-time. It provides the capability to collect, store, process, and visualize data as it is generated, allowing businesses to make informed decisions quickly based on up-to-date information.

**There are several widely used RTAP systems available in the market today. Here are three examples:**  
  
1. Apache Kafka: Apache Kafka is a distributed streaming platform that is often used as the backbone of real-time analytics systems. It provides a high-throughput, fault-tolerant messaging system that allows data to be ingested from multiple sources in real-time. Kafka's architecture enables horizontal scalability and fault tolerance, making it suitable for handling large volumes of data streams.  
  
2. Apache Flink: Apache Flink is an open-source stream processing framework that supports both batch and stream processing. It provides a unified programming model for processing real-time data streams and batch data sets. Flink's key features include event time processing, exactly-once semantics, stateful computations, and support for various data sources and sinks.  
  
3. Google Cloud Dataflow: Google Cloud Dataflow is a fully managed service for building real-time analytics pipelines. It allows users to define data processing workflows using a simple programming model and automatically scales the infrastructure based on the workload. Dataflow supports both batch and stream processing and integrates with other Google Cloud services such as BigQuery and Pub/Sub.

**DWDM-MST-II**

1. **Deference b/w classification and prediction. -2**

**Ans:**

| Aspect | Classification | Prediction |
| --- | --- | --- |
| Definition | Categorizing data into predefined classes or categories. | Estimating or forecasting a numerical outcome based on historical data. |
| Objective | Developing a model to assign new instances to specific classes. | Building a model to make informed predictions about future trends or events. |
| Output | Discrete class labels. | Numerical values for continuous variables. |
| Examples | Spam detection, image recognition, sentiment analysis, disease diagnosis. | Stock price forecasting, sales prediction, demand estimation. |

1. **Different type of clustering methods.**

Clustering is a fundamental technique in unsupervised machine learning that involves grouping a set of data points or objects based on their inherent characteristics or similarities. The goal of clustering is to identify patterns, similarities, or associations within a dataset without any predefined labels or target values. This process helps in understanding the natural grouping or structure of the data, enabling insights into the underlying relationships and distributions within the dataset.

Clustering algorithms partition the data into subsets, known as clusters, with each cluster ideally containing data points that are more similar to each other compared to data points in other clusters. The clusters can be formed based on various metrics, such as distance, density, or connectivity, depending on the specific algorithm and the nature of the data.

Some popular clustering algorithms include:

* **K-means Clustering:** A partitioning method that divides data points into K clusters, where each data point belongs to the cluster with the nearest mean.
* **Hierarchical Clustering**: An approach that creates a hierarchy of clusters, either by recursively merging or splitting clusters based on the distance between data points.
* **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** A density-based algorithm that identifies clusters as areas of high density separated by areas of low density.

1. **Accuracy, precision, true positive and false positive.**

**Ans: Accuracy**: Accuracy is a metric that measures the overall correctness of a model by calculating the ratio of the total number of correct predictions to the total number of predictions. In the context of binary classification, it can be represented as:

Accuracy=Number of Correct PredictionsTotal Number of PredictionsAccuracy=Total Number of PredictionsNumber of Correct Predictions​

**Precision**: Precision is a metric that assesses the accuracy of the positive predictions made by a model. It measures the ratio of true positive predictions to the total number of positive predictions made by the model. Precision is particularly useful when the cost of false positives is high. It can be expressed as:

Precision=True PositivesTrue Positives + False PositivesPrecision=True Positives + False PositivesTrue Positives​

**True Positive (TP)**: True positive refers to the number of correctly predicted positive instances by the model.

**False Positive (FP)**: False positive indicates the number of instances that were predicted as positive by the model but were actually negative.

1. **Application of clustering analysis.**

**Ans: Image Segmentation and Object Recognition**: In image analysis, clustering techniques are used for image segmentation, which involves dividing an image into meaningful segments. This process aids in object recognition, image compression, and various other computer vision applications.

**Anomaly Detection**: Clustering analysis can be employed for anomaly detection in various domains, such as fraud detection in financial transactions, network intrusion detection in cybersecurity, or identifying abnormalities in healthcare data. Clustering helps in identifying data points that deviate significantly from the normal behaviour.

**Recommendation Systems**: Clustering algorithms are used in recommendation systems to group users or items with similar characteristics. By identifying patterns in user preferences or item attributes, recommendation systems can suggest relevant products, services, or content to users, enhancing their overall experience.

1. **Hierarchical clustering.**

**Ans:** Hierarchical clustering is a type of clustering method that groups data points into clusters based on their similarities and differences. It is a hierarchical method, meaning that the clusters are constructed in a tree-like structure, with each cluster being divided into smaller sub-clusters. The algorithm starts by treating each data point as its own cluster, and then iteratively merges the closest clusters until only a single cluster remains.

1. **Search Engines its functionality with example. – 4**

**Ans:** A search engine is a software program or platform that retrieves and displays data or information from a database or the internet in response to a user's query or search request. The primary function of a search engine is to search for and retrieve information from a vast database of web pages, articles, documents, and other types of digital content.

**Examples of search engines include Google, Bing, and DuckDuckGo. Here's an example of how each of these search engines functions:**  
**\* Google:** When you enter a search query into Google, the search engine's algorithms analyze the query and use its vast index to quickly locate relevant web pages and other content. Google then ranks and sorts the results based on their relevance and popularity, displaying the most useful and accurate results at the top of the list. Google also offers advanced features such as spell checking, suggested search queries, and refinement tools that allow users to narrow their search results.  
**\* Bing:** Bing is another popular search engine that uses a similar approach to Google. It crawls the internet for new and updated content, stores the information in an index, and uses algorithms to rank and sort the results based on their relevance and popularity. Bing also offers advanced features such as spell checking and refinement tools.  
**\* DuckDuckGo:** DuckDuckGo is a privacy-focused search engine that emphasizes user privacy and avoids personalized search results. Instead of tracking user behaviour and collecting personal data, DuckDuckGo uses a traditional search algorithm that ranks and sorts results based on their relevance and popularity.

1. **Navie Byes and K-means algorithm.**

**Ans:** Naive Bayes Algorithm: Naive Bayes is a probabilistic machine learning algorithm used for classification tasks. It is based on Bayes' theorem, which calculates the probability of a hypothesis being true given the evidence. The "naive" assumption in Naive Bayes is that the features are conditionally independent of each other, simplifying the computation of probabilities.

**Functionalit**y: Naive Bayes calculates the probability of each class for a given set of features and selects the class with the highest probability as the predicted class for the input data.

**Application:** It is commonly used in text classification, spam filtering, sentiment analysis, and recommendation systems.

***K-means:***

The k-means algorithm is a popular unsupervised machine learning algorithm used for clustering analysis. It aims to partition a given dataset into k clusters, where each data point belongs to the cluster with the nearest mean value. The algorithm iteratively assigns data points to clusters and updates the cluster centres until convergence is achieved. K-means has various use cases across different domains due to its simplicity and efficiency.  
1. Image Compression:  
One of the primary applications of k-means is image compression. Images often contain a large number of colors, but not all colors are equally important for human perception. By applying k-means clustering on the color space of an image, we can reduce the number of colors used while preserving visual quality. The algorithm groups similar colors together and replaces them with the mean color value of each cluster. This reduces the storage space required for representing images without significant loss in visual fidelity.  
  
2. Customer Segmentation:  
K-means clustering is widely used in marketing and customer analytics for segmenting customers based on their behaviour, preferences, or demographics. By analyzing customer data such as purchase history, browsing patterns, or demographic information, businesses can identify distinct groups of customers with similar characteristics. These segments can then be targeted with personalized marketing campaigns or product recommendations, leading to improved customer satisfaction and higher conversion rates.  
  
3. Anomaly Detection:  
K-means can also be utilized for anomaly detection in various domains such as network security, fraud detection, or manufacturing quality control. By clustering normal instances of a system or process, any data point that significantly deviates from its assigned cluster can be considered an anomaly or outlier. This approach helps in identifying unusual patterns or behaviours that may indicate potential security breaches, fraudulent activities, or manufacturing defects.  
  
5. Recommendation Systems:  
K-means clustering can be used in recommendation systems to group users or items based on their preferences or attributes. By clustering users with similar tastes, personalized recommendations can be generated for individuals within the same cluster. Similarly, clustering items with similar features allows for recommending similar products or content to users based on their preferences.

**MST-II-NSC**

1. **AES, DES, RSA. – 2**

**AES (Advanced Encryption Standard):** AES is a symmetric encryption algorithm that is widely used to secure sensitive data. It was selected by the U.S. National Institute of Standards and Technology (NIST) to replace the older Data Encryption Standard (DES). AES operates on fixed block sizes of 128 bits and supports key sizes of 128, 192, or 256 bits. It is known for its efficiency and security, and it's commonly used for securing data in various applications, including securing data transmissions, data storage, and more.

**DES (Data Encryption Standard):** DES is an older symmetric encryption algorithm that was widely used in the 1970s and 1980s. It operates on 64-bit blocks and uses a 56-bit key. However, due to advances in computing power, DES is now considered relatively insecure and has been largely replaced by more robust algorithms like AES.

**RSA:** RSA is an asymmetric cryptographic algorithm named after its inventors, Ron Rivest, Adi Shamir, and Leonard Adleman. It is commonly used for secure data transmission and digital signatures. RSA relies on the practical difficulty of factoring the product of two large prime numbers, the factoring problem. The security of RSA is based on the practical difficulty of factoring the product of two large prime numbers, the factoring problem.

1. **Hash function, MD, MD5, secure hash, digest algo, Authentication function, digital signatures, key management. – 4**

**Ans:**

**Hash Function:**

A hash function is a mathematical function that takes an input (or 'message') and returns a fixed-size string of bytes, which is typically a hash value. Hash functions are used in various fields, including cryptography and data integrity verification.

**There are different types of hash functions, some of which include:**

**Cryptographic hash functions:** These are designed to be secure for use in cryptography and are typically used to verify the integrity of data. They have properties such as collision resistance, which means it is computationally infeasible to find two different inputs that produce the same hash value. Examples of cryptographic hash functions include SHA-256, SHA-3, and BLAKE2.

**Non-cryptographic hash functions:** These are used for data retrieval and data structure optimization. They are not designed for security purposes and may not possess the same level of collision resistance as cryptographic hash functions. Examples of non-cryptographic hash functions include Murmur Hash, City Hash, and Farm Hash.

**Checksums:** While not traditional cryptographic hash functions, checksums are also used to verify the integrity of data. They are used to detect errors in data transmission or storage. Examples of checksums include CRC (Cyclic Redundancy Check) and Adler-32.

Hash functions are fundamental in various applications, including data retrieval, data storage, password verification, and ensuring the integrity of transmitted data. Their properties and types determine their applicability in different contexts, ranging from security-sensitive operations to simple data validation tasks.

**MD (Message Digest):**

**MD (Message Digest):** MD refers to a series of cryptographic hash functions developed by Ronald Rivest. These functions generate fixed-size hash values from input data but have been replaced by more secure alternatives due to vulnerabilities.

Example: MD4, a widely used hash function, is no longer considered secure due to its vulnerability to collision attacks.

**Types of MD functions:**

* MD2: Produces a 128-bit hash value, now considered insecure.
* MD4: Also produces a 128-bit hash value, now considered insecure.
* MD5: Generates a 128-bit hash value, widely used but vulnerable to various attacks.

**MD5:**

MD5: MD5 is a widely used cryptographic hash function that produces a 128-bit (16-byte) hash value. It was commonly used for integrity checks and digital signatures but has vulnerabilities that make it unsuitable for certain cryptographic applications.

Example: MD5 is commonly used for verifying data integrity in various applications. However, due to its vulnerabilities, it has been largely replaced by more secure hash functions like SHA-2 and SHA-3.

| Advantages of MD5 Algorithm | Disadvantages of MD5 Algorithm |
| --- | --- |
| Faster and simple to understand | Generates the same hash function for different inputs |
| Generates a strong 16-byte password format | Provides poor security compared to SHA1 |
| Requires relatively low memory for integration | Considered an insecure algorithm |
| Easy and fast to generate a digest message | Not suitable as a symmetric or asymmetric algorithm |

**Secure Hash Algorithm (SHA):**

Secure Hash Algorithm (SHA): SHA is a family of cryptographic hash functions designed by the National Security Agency (NSA) and published by the National Institute of Standards and Technology (NIST). It is widely used in various security applications, including digital signatures and data integrity verification.

Example: SHA-1 is one of the members of the SHA family, producing a 160-bit hash value. It has been widely used but is now considered vulnerable to collision attacks.

**The various types of the Secure Hash Algorithm (SHA) include:**

**SHA-1 (Secure Hash Algorithm 1):** It produces a 160-bit hash value and was commonly used for various cryptographic applications. However, due to vulnerabilities and the advent of more secure alternatives, its usage has significantly declined.

**SHA-2 (Secure Hash Algorithm 2):** SHA-2 is a family of cryptographic hash functions that includes various hash sizes, such as SHA-224, SHA-256, SHA-384, and SHA-512. These algorithms are widely used for data integrity verification, digital signatures, and other security applications.

**SHA-3 (Secure Hash Algorithm 3):** SHA-3 is the latest member of the SHA family, providing an alternative approach to cryptographic hashing compared to SHA-2. It was developed as a response to potential security concerns and is designed to provide additional security against potential attacks.

**Digest Algorithm:**

Digest Algorithm: A digest algorithm is a cryptographic hash function used to produce a fixed-size hash value from input data. It is commonly used for data integrity verification and secure communication.

Example: The RIPEMD (RACE Integrity Primitives Evaluation Message Digest) family of hash functions, including RIPEMD-160, is used in various security protocols for data integrity verification and secure communication.

**Here are some commonly used types of digest algorithms:**

**MD5 (Message Digest Algorithm 5):** MD5 is a widely used digest algorithm that generates a 128-bit (16-byte) hash value. It was commonly used for integrity checks and digital signatures. However, due to vulnerabilities such as collision attacks, it is no longer considered secure for cryptographic use.

**SHA-1 (Secure Hash Algorithm 1):** SHA-1 is a commonly used digest algorithm that produces a 160-bit (20-byte) hash value. While it was widely used for data integrity and digital signatures, its security has been compromised, and it is no longer considered secure for cryptographic purposes.

**SHA-256, SHA-384, and SHA-512:** These are part of the SHA-2 (Secure Hash Algorithm 2) family, with hash sizes of 256, 384, and 512 bits, respectively. They are widely used in various security applications and are considered more secure than their predecessors, such as MD5 and SHA-1.

**SHA-3 (Secure Hash Algorithm 3):** SHA-3 is the latest member of the Secure Hash Algorithm family. It was developed as a response to potential security concerns and is designed to provide additional security against potential attacks, offering an alternative to SHA-2.

**Authentication Function:**

Authentication Function: An authentication function is used to verify the identity of a message sender or receiver, ensuring that they are who they claim to be.

Example: HMAC (Hash-based Message Authentication Code) is widely used for message authentication, providing data integrity and authentication for messages using cryptographic hash functions.

**Authentication functions can take various forms, including:**

**Password-based Authentication:** This type of authentication involves users providing a password or passphrase to verify their identity. The system compares the provided password with the stored one and grants access if they match.

**Biometric Authentication:** Biometric authentication utilizes unique biological characteristics such as fingerprints, iris patterns, or facial features to verify an individual's identity.

**Multi-factor Authentication (MFA):** MFA requires users to provide two or more different authentication factors to gain access. These factors can include something they know (password), something they have (a security token or mobile device), or something they are (biometric data).

**Digital Signatures:**

Digital Signatures: Digital signatures are cryptographic techniques used to provide authenticity and integrity to digital documents or messages, ensuring that the contents have not been altered and that the sender cannot deny sending the message.

Example: DSA (Digital Signature Algorithm) is commonly used for digital signatures, ensuring the authenticity and integrity of digital documents and messages.

**Here are some common types of digital signatures:**

**RSA Digital Signature:** RSA (Rivest-Shamir-Adleman) is a widely used algorithm for digital signatures. It uses asymmetric key cryptography, with a private key to sign the message and a corresponding public key to verify the signature.

**DSA (Digital Signature Algorithm):** DSA is a United States government standard for digital signatures. It uses the Digital Signature Algorithm to provide data integrity and authentication, ensuring that the message has not been tampered with during transmission.

**Key Management:**

Key management refers to the processes and procedures involved in generating, storing, distributing, using, and replacing cryptographic keys. It is a crucial aspect of cryptographic systems and protocols, ensuring the security and integrity of data transmissions and communications. Effective key management is essential for maintaining the confidentiality, integrity, and availability of sensitive information in various applications.

**There are different types of key management strategies and techniques, including:**

**Symmetric Key Management:** Symmetric key management involves the secure distribution and storage of symmetric keys used for encryption and decryption. It includes processes such as key generation, key storage, key distribution, and key rotation to ensure the secure exchange of information between parties.

**Asymmetric Key Management:** Asymmetric key management deals with the secure generation, distribution, and storage of public and private key pairs used in public-key cryptography. It involves processes such as key pair generation, certificate management, and key revocation to establish secure communication channels and enable digital signatures and encryption.

**Key Exchange Protocols:** Key exchange protocols facilitate the secure exchange of cryptographic keys between parties over an insecure communication channel. Protocols like Diffie-Hellman key exchange and its variants enable secure key exchange without the need for pre-shared secrets.

1. **Firewalls, types of firewalls honeypot. – 8**

Firewalls are security systems designed to monitor and control network traffic, acting as a barrier between trusted internal networks and untrusted external networks, such as the internet. They examine incoming and outgoing network traffic and decide whether to allow or block specific traffic based on predetermined security rules. Firewalls help protect networks from unauthorized access, cyberattacks, and other security threats.

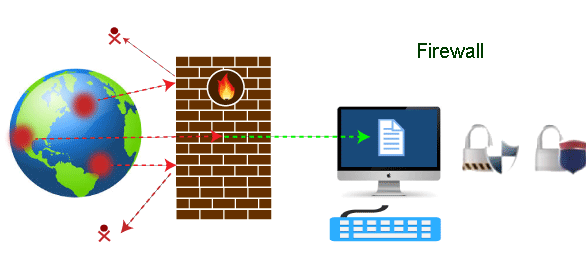
**There are several types of firewalls, including:**

**Packet-Filtering Firewalls:** Packet-filtering firewalls inspect individual packets of data as they pass through the firewall. They make decisions based on factors such as source and destination IP addresses, ports, and protocols. However, they provide basic security and may not offer more advanced filtering capabilities.

**Stateful Inspection Firewalls:** Stateful inspection firewalls monitor the state of active connections and make decisions based on the context of the traffic. They keep track of the state of network connections and use this information to make more informed decisions about which packets to allow or block.

**Proxy Firewalls:** Proxy firewalls act as an intermediary between internal and external networks. They receive network requests on behalf of clients and initiate the requests on their behalf. Proxy firewalls can provide additional security by hiding internal IP addresses and performing content filtering.

**Next-Generation Firewalls (NGFWs):** NGFWs combine traditional firewall functionality with additional features, such as intrusion detection and prevention, deep packet inspection, and application awareness. They offer more advanced security capabilities and can provide better protection against modern security threats.



**How does a firewall work?**

A firewall system analyzes network traffic based on pre-defined rules. It then filters the traffic and prevents any such traffic coming from unreliable or suspicious sources. It only allows incoming traffic that is configured to accept.

Typically, firewalls intercept network traffic at a computer's entry point, known as a port. Firewalls perform this task by allowing or blocking specific data packets (units of communication transferred over a digital network) based on pre-defined security rules. Incoming traffic is allowed only through trusted [IP](https://www.javatpoint.com/ip) addresses, or sources.

**Honeypot**

A **honeypot** is a security mechanism used to detect, deflect, or counteract unauthorized access to information systems. It consists of a computer, data, or network site that appears to be part of a network but is isolated and closely monitored to attract and trap potential intruders. Honeypots are designed to gather information about attackers' tactics, techniques, and tools and provide insights into potential security vulnerabilities.

**There are different types of honeypots, including:**

**Low-Interaction Honeypots:** Low-interaction honeypots simulate a limited set of services and protocols to attract potential attackers. They provide basic information about attempted attacks without exposing the entire system to risk.

**High-Interaction Honeypots:** High-interaction honeypots emulate real systems and services, allowing attackers to interact with a complete and fully functional environment. They capture more detailed information about attacker behavior but pose a higher risk as they provide a more realistic environment.

Honeypots serve as valuable tools for security researchers and organizations to study the techniques and methods used by attackers, identify potential vulnerabilities, and improve overall network security. However, they should be implemented carefully to minimize the risk of exposing sensitive information and resources to malicious actors.

**AWSN-MST-II**

1. **Difference between RNN and BRNN. -2**

**Ans:**

| Aspect | Recurrent Neural Network (RNN) | Bidirectional Recurrent Neural Network (BRNN) |
| --- | --- | --- |
| Information Flow | Processes input sequentially in one direction. | Processes input in both directions simultaneously. |
| Capturing Context | Limited to capturing context from past inputs only. | Capable of capturing context from both past and future inputs. |
| Use Cases | Suitable for tasks where understanding past context is sufficient. | Beneficial for tasks where both past and future contexts are vital. |
| Training | Can be easier to train compared to BRNN. | More challenging to train, especially on large datasets. |
| Computational Complexity | Relatively less complex compared to BRNN. | More complex and computationally expensive compared to RNN. |

1. **Pooling.**

**Ans:** pooling is an essential operation in deep learning that helps in reducing spatial dimensions, achieving translation invariance, and extracting relevant features. It has become a standard component in CNN architectures and plays a crucial role in various computer vision tasks. While it has its limitations, advancements in pooling techniques continue to improve its effectiveness and flexibility.

1. **DNN.**

**Ans:** Deep Neural Networks (DNNs) are a type of artificial neural network (ANN) that are widely used in the field of deep learning. DNNs are designed to mimic the structure and functionality of the human brain, allowing them to learn and make predictions from complex data.

1. **Application of RNN.**

Ans: **Speech Recognition and Synthesis:** RNNs play a significant role in speech-related tasks, including speech recognition, speech-to-text conversion, and voice synthesis.

**Natural Language Processing (NLP):** RNNs are extensively employed in NLP tasks such as language modelling, machine translation, sentiment analysis, text generation, and speech recognition.

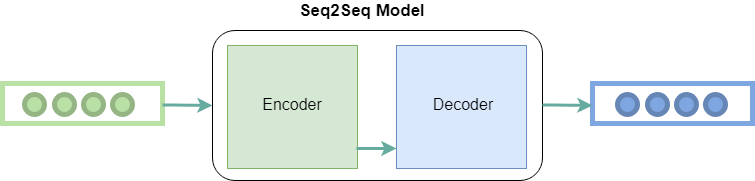
1. **Sequence 2 Sequence. – 4**

**Ans:** Sequence-to-sequence (Seq2Seq) is a deep learning architecture that has been widely used for various natural language processing (NLP) tasks, such as machine translation, text summarization, and speech recognition. It is a type of neural network model that can map an input sequence to an output sequence of potentially different lengths.

**The Seq2Seq model consists of two main components: an encoder and a decoder.**

**The encoder** takes the input sequence and converts it into a fixed-length vector representation called the context vector or the hidden state. This context vector captures the semantic meaning of the input sequence.

**The decoder** then takes this context vector as input and generates the output sequence step by step.



| Advantages | Disadvantages |
| --- | --- |
| Handling Variable-Length Input and Output | Complexity and Computation |
| Natural Language Processing (NLP) Applications | Data Dependency |
| End-to-End Learning | Overfitting and Generalization |
| Attention Mechanism | Inference Time Complexity |

1. **CNN.**

**Ans: Convolutional Neural Network** is one of the main categories to do image classification and image recognition in neural networks. Scene labelling, objects detections, and face recognition, etc., are some of the areas where convolutional neural networks are widely used.

CNN takes an image as input, which is classified and process under a certain category such as dog, cat, lion, tiger, etc. The computer sees an image as an array of pixels and depends on the resolution of the image. Based on image resolution, it will see as **h \* w \* d**, where h= height w= width and d= dimension. For example, An RGB image is **6 \* 6 \* 3** array of the matrix, and the grayscale image is **4 \* 4 \* 1** array of the matrix.

In CNN, each input image will pass through a sequence of convolution layers along with pooling, fully connected layers, filters (Also known as kernels). After that, we will apply the Soft-max function to classify an object with probabilistic values 0 and 1.

**Layers:**

**Convolutional Layers:** These layers apply learnable filters to local regions of the input data to extract essential features and patterns.

**Pooling Layers:** Pooling layers reduce the spatial dimensions of the feature maps by summarizing local information, making the network translation-invariant and reducing computational complexity.

**Fully Connected Layers:** Fully connected layers integrate the features learned by previous layers and make predictions or classifications based on the extracted features.

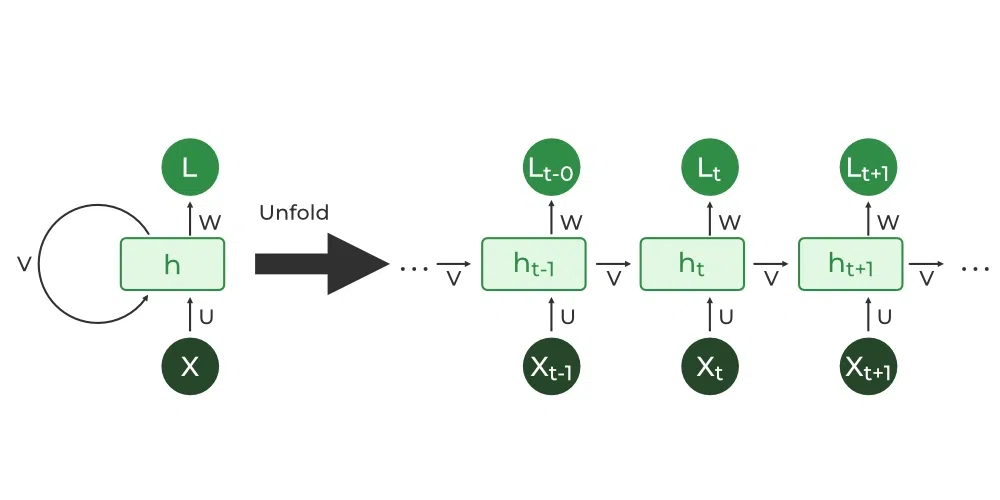
|  |
| --- |
|  |
| Advantages | Disadvantages |
| Automated Feature Learning | Complexity and Computational Resources |
| Spatial Hierarchical Representations | Large Data Requirements |
| State-of-the-Art Performance |  |

1. **RNN with example and advantages and disadvantages.-8s:**

**Ans:**

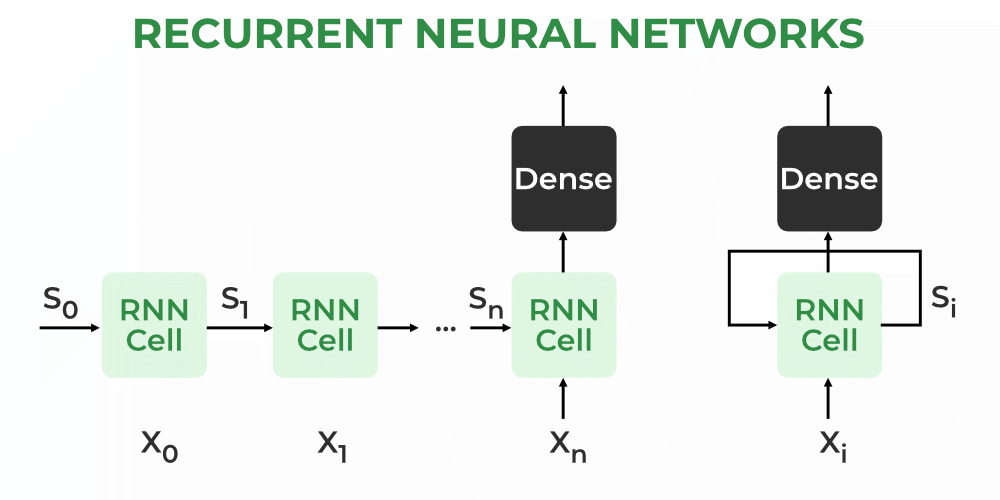
Recurrent Neural Network(RNN) is a type of [Neural Network](https://www.geeksforgeeks.org/tag/neural-network/) where the output from the previous step is fed as input to the current step.

The main and most important feature of RNN is its **Hidden state**, which remembers some information about a sequence.



**Architecture Of Recurrent Neural Network**

RNNs have the same input and output architecture as any other deep neural architecture. However, differences arise in the way information flows from input to output.



**How  RNN works**

The Recurrent Neural Network consists of multiple fixed activation function units, one for each time step. Each unit has an internal state which is called the hidden state of the unit. This hidden state signifies the past knowledge that the network currently holds at a given time step.

|  |
| --- |
|  |
| Advantages | Disadvantages |
| Remembers previous inputs through time | Gradient vanishing and exploding problems |
| Suitable for time series prediction | Training an RNN is a challenging task |
| Enables Long Short-Term Memory (LSTM) | Difficulty processing long sequences with certain activation functions |
| Can be combined with convolutional layers to extend effective pixel neighbourhood |  |

**Types Of RNN**

There are four types of RNNs based on the number of inputs and outputs in the network.

One to One

One to Many

Many to One

Many to Many

**One to One**

This type of RNN behaves the same as any simple Neural network it is also known as Vanilla Neural Network. In this Neural network, there is only one input and one output.

**One To Many**

In this type of RNN, there is one input and many outputs associated with it. One of the most used examples of this network is Image captioning where given an image we predict a sentence having Multiple words.

**Many to One**

In this type of network, many inputs are fed to the network at several states of the network generating only one output. This type of network is used in the problems like sentimental analysis. Where we give multiple words as input and predict only the sentiment of the sentence as output.

**Many to Many**

In this type of neural network, there are multiple inputs and multiple outputs corresponding to a problem. One Example of this Problem will be language translation. In language translation, we provide multiple words from one language as input and predict multiple words from the second language as output.

1. **Proactive (Table-Driven) Protocols:**Proactive protocols, also known as table-driven protocols, are a type of network protocol used in computer networks to facilitate efficient routing of data packets. These protocols rely on maintaining and periodically updating routing tables at each node in the network. The routing tables contain information about the network topology, including the available paths and their associated costs.
2. **. Reactive (On-Demand) Protocols:**Reactive protocols, or on-demand protocols, are communication protocols that establish routes only when there is a need for communication. They help reduce overhead and conserve network resources by avoiding periodic control message exchanges.
3. **Hybrid Protocols:**  
   Hybrid protocols are a combination of two or more different communication protocols used to achieve a specific goal or provide a particular service. These protocols can be used in various applications, such as network communication, data transfer, and security. The main advantage of hybrid protocols is that they can leverage the strengths of multiple protocols to provide a more robust and reliable solution.
4. **GPP.**

**Ans:** GPPs are responsible for executing various tasks related to data processing, communication, and network management. They provide the necessary computational power to perform complex operations on the collected data and make intelligent decisions based on the application requirements. GPPs are typically more powerful than the microcontrollers used in sensor nodes, enabling them to handle more computationally intensive tasks.

1. **Goals of TCP solution as compare to Adhoc.**

**Ans:** The Transmission Control Protocol (TCP) and ad hoc networks are two different solutions with distinct goals. TCP is a widely used protocol in computer networking that provides reliable, ordered, and error-checked delivery of data packets over an IP network. On the other hand, ad hoc networks are decentralized wireless networks where devices communicate directly with each other without the need for a centralized infrastructure.

**The goals of TCP as compared to ad hoc networks can be summarized as follows:**

**Reliability:** One of the primary goals of TCP is to ensure reliable data transmission. It achieves this by using various mechanisms such as acknowledgments, retransmissions, and flow control. TCP guarantees that data sent from one device will be received correctly by the destination device.

In contrast, ad hoc networks may have limited reliability due to the dynamic nature of the network topology and the absence of a centralized control mechanism.

**Ordering:** TCP ensures that data packets are delivered in the same order they were sent. This is important for applications that require sequential delivery of data, such as streaming media or file transfers. Ad hoc networks, on the other hand, may not guarantee strict ordering of packets due to the decentralized nature of communication between devices.

**Congestion Control:** TCP employs congestion control to prevent network congestion and ensure equitable resource sharing. It adjusts the sending rate dynamically based on network conditions.

Ad hoc networks face challenges in implementing congestion control due to the lack of centralized control and dynamic topology. Maintaining smooth data flow without overwhelming the network is complex in ad hoc environments.

1. **Challenges WSN, enabling technology for WSN.**

**Ans:** Wireless Sensor Networks (WSNs) are a type of network that consists of numerous autonomous sensor nodes deployed in a specific area to monitor physical or environmental conditions. These networks have gained significant attention due to their potential applications in various fields, including environmental monitoring, healthcare, industrial automation, and smart cities. However, WSNs also face several challenges that need to be addressed for their successful deployment and operation. In this response, we will discuss some of the key challenges faced by WSNs and the enabling technologies that can help overcome these challenges.

**Challenges:**

* Energy Consumption and Lifetime
* Reliability and Availability
* Security and Privacy
* Scalability and Interoperability
* Data Processing and Management

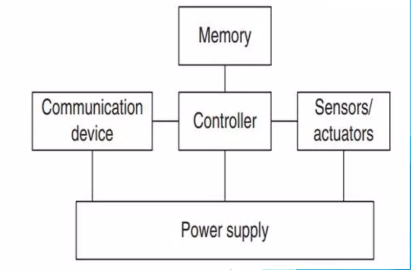
**Enabling Technologies:**

* Low-Power Wireless Communication Protocols
* Energy-Efficient Sensing and Actuation
* Energy Harvesting and Storage
* Redundancy and Fault Tolerance
* Security and Privacy Protocols

1. **Single Node Architecture.**

**Ans:** Single node architecture refers to a system or network architecture that consists of a single node or server. In this architecture, all the processing, storage, and communication tasks are performed by a single machine. The term “node” refers to a physical or virtual entity that can execute tasks and communicate with other nodes in a network.

**Architecture:** In a single node architecture, all the software components, databases, and services are installed and executed on a single machine.

****

**Component:**

**Controller:** It I the central processing unit of the node it collects data from sensor and process and decides when and where to send it capable of execution arbitrary code.

**Memory:** Memory are used to store program and intermediate data; usually, different type of memory are used for programs and data.

**Sensor:** The actual interface to the physical world: devices that can observer or control physical parameters of the environment.

**Actuators**: These are for wireless sensor networks that convers Electrical signal into physical phenomenon.

**Communication Device**: Turning nodes into a network requires a device for sending and receiving information over wireless channel, here it is used Radio frequencies.

**Power Supply**: As usually no tethered power supply is available, some forms of batteries are necessary to provide energy, sometimes, some form of recharging by obtaining energy from environment is available as well (e.g., solar cells)

**Advantages Vs Disadvantages:**

| **Advantages** | **Disadvantages** |
| --- | --- |
| Simplicity | Scalability |
| Cost-effective | Single point of failure |
| Performance | Limited resource utilization |

Example: An example of a single node architecture is a personal computer (PC) or a standalone server. In these systems, all the processing, storage, and communication tasks are performed by a single machine. The PC or server runs an operating system and various applications that are executed locally on the machine. The resources of the machine, such as CPU, memory, and storage, are dedicated to serving the needs of that single node.