Answer Submitted.
X

(https://swayam.gov.in)

(https://swayam.gov.in/nc_details/NPTEL)

gireesh218@gmail.com

NPTEL (https://swayam.gov.in/explorer?ncCode=NPTEL) » Big Data Computing (course)

Announcements (announcements) About the Course (preview) Q&A (forum) Progress (student/home) Mentor (student/mentor)

Review Assignment (assignment_review) Course Recommendations New! (/course_recommendations)

Course outline

About NPTEL ()

How does an NPTEL online course work? ()

Week-0 ()

○ **Practice: Week 0: Assignment 0**

Week 0: Assignment 0

Assignment not submitted

1) What defines Big Data? ***1 point***

● Volume, Variety, Velocity
○ Veracity, Velocity, Value
○ Volume, Veracity, Value
○ Value, Versatility, Volume

Yes, the answer is correct.

(assessment?name=130)

Week-1 ()

Week-2 ()

Week-3 ()

Week-4 ()

Week-5 ()

Week-6 ()

Week-7 ()

Text Transcripts ()

DOWNLOAD VIDEOS ()

Books ()

Score: 1

Accepted Answers:
Volume, Variety, Velocity

2) Which technology is commonly used for processing and analyzing Big Data? ***1 point***

● Hadoop
○ SQL
○ Python
○ Excel

Yes, the answer is correct.
Score: 1

Accepted Answers:
Hadoop

3) Which of the following is a challenge associated with Big Data? ***1 point***

○ Low storage requirements
○ Limited data sources
● Slow data processing
○ Predictable data patterns

Yes, the answer is correct.
Score: 1

Accepted Answers:
Slow data processing

4) Which programming language is commonly used in Hadoop development? ***1 point***

● Java
○ Python
○ C++
○ Ruby

Yes, the answer is correct.
Score: 1

Accepted Answers:
Java

5) What is the primary purpose of Hadoop's HDFS? **1 point**

○ Data visualization
● Data storage
○ Data querying
○ Data modeling

Yes, the answer is correct.
Score: 1

Accepted Answers:
Data storage

6) Which component of Hadoop is responsible for job scheduling and resource management? **1 point**

● YARN
○ HDFS
○ MapReduce
○ Pig

Yes, the answer is correct.
Score: 1

Accepted Answers:
YARN

7) What is Apache Zookeeper primarily used for in Big Data ecosystems? **1 point**

○ Data storage
○ Data processing
● Configuration management

○ Data visualization

Yes, the answer is correct.
Score: 1

Accepted Answers:
Configuration management

8) What is the default block size in HDFS? **1 point**

● 128 MB
○ 256 MB
○ 64 MB
○ 512 MB

Yes, the answer is correct.
Score: 1

Accepted Answers:
128 MB

9) CAP theorem states that a distributed system cannot simultaneously guarantee? **1 point**

● Consistency, Accessibility, Partition tolerance
○ Consistency, Atomicity, Partition tolerance
○ Consistency, Atomicity, Availability
○ Consistency, Availability, Reliability

Yes, the answer is correct.
Score: 1

Accepted Answers:
Consistency, Accessibility, Partition tolerance

10) Which of the following is NOT a role of Apache Zookeeper? **1 point**

○ Data storage
○ Data processing
● Configuration management

○ Data visualization

Yes, the answer is correct.
Score: 1

Accepted Answers:
Configuration management

Check Answers and Submit

Your score is: 10/10

Assignment 1

1. Which of the following best describes the concept of 'Big Data'?
    a. Data that is physically large in size
    b. Data that is collected from multiple sources and is of high variety, volume, and velocity
    c. Data that requires specialized hardware for storage
    d. Data that is highly structured and easily analyzable

Ans- Big Data is characterized by the "Three Vs": variety (different types of data), volume (large amounts of data), and velocity (speed at which data is generated and processed). This definition captures the essence of Big Data, distinguishing it from merely large or structured datasets.

2. Which technology is commonly used for processing and analyzing Big Data in distributed computing environments?
    a. MySQL
    b. Hadoop
    c. Excel
    d. SQLite

Ans- Hadoop is a widely-used framework designed for processing and analyzing large datasets in distributed computing environments. It provides a scalable and fault-tolerant way to handle Big Data, unlike MySQL, Excel, or SQLite, which are not typically used for large-scale distributed processing.

3. What is a primary limitation of traditional RDBMS when dealing with Big Data?
    a. They cannot handle structured data
    b. They are too expensive to implement
    c. They struggle with scaling to manage very large datasets
    d. They are not capable of performing complex queries

Ans- Traditional Relational Database Management Systems (RDBMS) often face challenges with scalability when handling Big Data, primarily due to their limited ability to distribute data across multiple nodes. They are not inherently designed for the scale required by Big Data.

4. Which component of Hadoop is responsible for distributed storage?

a. YARN
b. HDFS
c. MapReduce
d. Pig

Ans- The Hadoop Distributed File System (HDFS) is the component responsible for storing data across a distributed cluster, providing redundancy and fault tolerance. YARN is for resource management, MapReduce is a processing framework, and Pig is a high-level data flow language.

5. Which Hadoop ecosystem tool is primarily used for querying and analyzing large datasets stored in Hadoop's distributed storage?

a. HBase
b. Hive
c. Kafka
d. Sqoop

Ans- Hive is a data warehousing and SQL-like query language tool used to query and analyze large datasets in Hadoop. HBase is a NoSQL database, Kafka is a messaging system, and Sqoop is used for data transfer between Hadoop and relational databases.

6. Which YARN component is responsible for coordinating the execution of tasks within containers on individual nodes in a Hadoop cluster?

a. NodeManager
b. ResourceManager
c. ApplicationMaster
d. DataNode

Ans- NodeManager is the YARN component responsible for managing resources and monitoring the execution of tasks on individual nodes. ResourceManager manages overall cluster resources, ApplicationMaster handles application-specific resource requests, and DataNode is part of HDFS.

7. What is the primary advantage of using Apache Spark over traditional MapReduce for data processing?

a. Better fault tolerance
b. Lower hardware requirements
c. Real-time data processing
d. Faster data processing

Ans- Apache Spark provides faster data processing compared to traditional MapReduce due to its in-memory processing capabilities, which reduce the need for disk I/O operations. This leads to significant performance improvements for iterative algorithms and complex data processing tasks.

8. What is Apache Spark Streaming primarily used for?

a. Real-time data visualization
b. Batch processing of large datasets
c. Real-time stream processing
d. Data storage and retrieval

Ans- Apache Spark Streaming is designed for real-time stream processing, enabling the analysis of live data streams. It is not used for batch processing, real-time visualization, or data storage and retrieval.

9. Which operation in Apache Spark GraphX is used to perform triangle counting on a graph?

a. `connectedComponents`
b. `triangleCount`
c. `shortestPaths`
d. `pageRank`

And-The `triangleCount` operation in Apache Spark GraphX is used to count the number of triangles in a graph, which helps in analyzing the structure and connectivity of the graph.

10. Which component in Hadoop is responsible for executing tasks on individual nodes and reporting back to the JobTracker?

a. HDFS Namenode
b. TaskTracker
c. YARN ResourceManager
d. DataNode

Ans- The TaskTracker is responsible for executing MapReduce tasks on individual nodes and reporting the progress and status back to the JobTracker. The HDFS Namenode manages the file system namespace, the YARN ResourceManager allocates resources, and DataNode stores the actual data.

Assignmet -2

1. Which statement best describes the data storage model used by HBase?
    a. Key-value pairs
    b. Document-oriented
    c. Encryption
    d. Relational tables

Ans-

Option a: Key-value pairs - Correct. HBase is a NoSQL database that stores data in a key-value format. Each row is identified by a unique key, and columns within a row are organized into column families.

Option b: Document-oriented - Incorrect. Document-oriented databases like MongoDB store data as self-contained documents, while HBase uses a key-value model.

Option c: Encryption- Incorrect.

Option d: Relational tables - Incorrect. Relational databases use structured tables with rows and columns, while HBase is a NoSQL database with a flexible schema.

2. What is Apache Avro primarily used for in the context of Big Data?
    a. Real-time data streaming
    b. Data serialization
    c. Machine learning
    d. Database management

Ans-

Option a: Real-time data streaming - Incorrect. While Avro can be used in streaming applications, its primary focus is data serialization.

Option b: Data serialization - Correct. Avro is a data serialization format that efficiently encodes data structures for storage and transmission.

Option c: Machine learning - Incorrect. Avro can be used to store data for machine learning models, but its core functionality is data serialization.
Option d: Database management - Incorrect. Avro is not a database management system, but a format for storing data.

Explanation-
Apache Avro is a framework for data serialization. It provides a compact, fast, and efficient way to serialize and deserialize data, making it suitable for communication between different systems or for persisting data in a binary format. Avro is commonly used in Big Data applications for serialization of data in a way that supports schema evolution and provides interoperability across various programming languages.

3.Which component in HDFS is responsible for storing actual data blocks on the DataNodes?

a. NameNode
b. DataNode
c. Secondary NameNode
d. ResourceManager

Ans-option a: NameNode - Incorrect. The NameNode manages metadata about the file system, such as block locations and file permissions.
option b: DataNode - Correct. DataNodes are the physical storage units in HDFS that store data blocks.
option c: Secondary NameNode - Incorrect. The Secondary NameNode is a backup of the NameNode, not for data storage.
option d: ResourceManager - Incorrect. The ResourceManager is part of YARN, responsible for resource management in Hadoop, not data storage.
Explanation-

In the Hadoop Distributed File System (HDFS), DataNodes are responsible for storing the actual data blocks. Each DataNode manages the storage of data blocks and periodically sends heartbeat signals and block reports to the NameNode to confirm its status and the health of the data blocks it stores.

4.Which feature of HDFS ensures fault tolerance by replicating data blocks across multiple DataNodes?

a. Partitioning
b. Compression
c. Replication
d. Encryption

Ans- option a: partitioning - Incorrect. Partitioning divides data into smaller chunks for processing, not for fault tolerance.
option b: compression - Incorrect. Compression reduces data size, but doesn't provide redundancy.
option c: replication - Correct. HDFS replicates data blocks across multiple DataNodes to ensure data availability in case of node failures.
option d: encryption - Incorrect. Encryption protects data confidentiality, not availability.

Explanation-
HDFS achieves fault tolerance through the replication of data blocks. Each data block is replicated across multiple DataNodes, which helps in ensuring data availability and reliability even in the event of hardware failures. By default, each block is replicated three times across different nodes to safeguard against data loss.

5.Which component in MapReduce is responsible for sorting and grouping the intermediate key-value pairs before passing them to the Reducer?

a. Mapper
b. Reducer
c. Partitioner
d. Combiner

Ans- option a: Mapper - Incorrect. The Mapper generates key-value pairs, but doesn't perform sorting or grouping.
option b: Reducer - Incorrect. The Reducer processes the grouped key-value pairs, but doesn't perform the initial sorting and grouping.
option c: Partitioner - Correct. The Partitioner determines which Reducer will process a specific key-value pair.

option d: Combiner - Incorrect. The Combiner is an optional optimization that can reduce data volume before sending it to the Reducer, but it doesn't perform sorting and grouping.

Explanation-

In the MapReduce framework, the Partitioner is responsible for distributing the intermediate key-value pairs generated by the Mapper to the appropriate Reducer tasks. It handles the sorting and grouping of these pairs to ensure that all values for a given key are sent to the same Reducer. The sorting and grouping happen as part of the shuffle and sort phase of the MapReduce process.

6.What is the default replication factor in Hadoop Distributed File System (HDFS)?

a. 1
b. 2
c. 3
d. 4

Ans- option a: 1 - Incorrect. A replication factor of 1 would offer no fault tolerance.
option b: 2 - Incorrect. A replication factor of 2 provides some fault tolerance, but 3 is the default.
option c: 3 - Correct. The default replication factor in HDFS is 3, providing a balance between fault tolerance and storage efficiency.
option d: 4 - Incorrect. A replication factor of 4 would increase storage overhead without significantly improving fault tolerance.

Explanation-

The default replication factor in HDFS is 3. This means that each data block is replicated three times across different DataNodes. This default replication factor strikes a balance between data redundancy and storage overhead, providing fault tolerance and high availability for the data.

7.In a MapReduce job, what is the role of the Reducer?

a. Sorting input data
b. Transforming intermediate data
c. Aggregating results
d. Splitting input data

Ans- option a: sorting input data - Incorrect. The Mapper and Partitioner handle data sorting and distribution.
option b: transforming intermediate data - Correct. The Reducer can transform intermediate data based on the key-value pairs it receives.
option c: aggregating results - Correct. The Reducer is often used to aggregate values based on the key.
option d: splitting input data - Incorrect. The input data is split into blocks by the InputFormat.
Explanation-

The Reducer in a MapReduce job is responsible for aggregating the intermediate data produced by the Mapper. It takes the sorted and grouped key-value pairs from the shuffle and sort phase and performs a reduction operation, which might involve summing up values, calculating averages, or other forms of aggregation depending on the specific job requirements.

8.Which task can be efficiently parallelized using MapReduce?
a. Real-time sensor data processing
b. Single-row database queries
c. Image rendering
d. Log file analysis

Ans- option a: real-time sensor data processing - Incorrect. MapReduce is better suited for batch processing than real-time processing.
option b: single-row database queries - Incorrect. Single-row database queries are typically handled by relational databases.
option c: image rendering - Incorrect. Image rendering often requires specialized hardware and algorithms.
option d: log file analysis - Correct. Log file analysis involves processing large amounts of data, making it a good candidate for MapReduce.

Explanation-
MapReduce is particularly well-suited for tasks that can be parallelized across a large number of independent data chunks. Log file analysis is an example of such a task, as log files can be split into segments that can be processed in parallel. Each Mapper processes a chunk of log data to extract relevant information, and the Reducer aggregates and processes the results.

9.Which MapReduce application involves counting the occurrence of words in a large corpus of text?

a. PageRank algorithm
b. K-means clustering
c. Word count
d. Recommender system

Ans- option a: PageRank algorithm - Incorrect. PageRank is used for ranking web pages.
option b: K-means clustering - Incorrect. K-means clustering is used for grouping data points.
option c: word count - Correct. A word count application counts the frequency of words in a text corpus.
option d: recommender system - Incorrect. Recommender systems use collaborative filtering or content-based approaches.

Explanation-
The Word Count application is a classic example of a MapReduce job. It involves counting the frequency of each word in a large corpus of text. The Mapper extracts words from the text and emits them as key-value pairs with a count of 1. The Reducer then sums up these counts for each unique word to produce the final word count results.

10.What does reversing a web link graph typically involve?

a. Removing dead links from the graph
b. Inverting the direction of edges
c. Adding new links to the graph
d. Sorting links based on page rank

Ans- option a: removing dead links from the graph - Incorrect. Removing dead links is a different task.
option b: inverting the direction of edges - Correct. Reversing a web link graph means changing the direction of links, creating a graph where pages are pointed to instead of pointing to others.
option c: adding new links to the graph - Incorrect. Reversing doesn't involve adding new links.
option d: sorting links based on page rank - Incorrect. Sorting links is a different operation.

Explanation-
Reversing a web link graph involves inverting the direction of the edges between nodes (web pages). In a web link graph, each directed edge represents a hyperlink from one page to another. Reversing the graph means changing the direction of these links, so a link from Page A to Page B becomes a link from Page B to Page A. This is useful for various analyses, such as computing PageRank in a different context or understanding link relationships from a different perspective.

assignment 3

1. Which abstraction in Apache Spark allows for parallel execution and distributed data processing?
    a. DataFrame
    b. RDD (Resilient Distributed Dataset)
    c. Dataset
    d. Spark SQL

Option a: DataFrame - Incorrect. DataFrames provide a higher-level API for structured data, but they are not the fundamental abstraction for parallel execution.
Option b: RDD (Resilient Distributed Dataset) - **Correct**. RDDs are the fundamental abstraction in Spark for distributed, fault-tolerant, and parallel processing of large datasets.
Option c: Dataset - Incorrect. Datasets are a more recent addition to Spark, combining the benefits of RDDs and DataFrames. However, RDDs are the core concept for parallel execution.
Option d: Spark SQL - Incorrect. Spark SQL is a SQL engine built on top of Spark, providing SQL-like capabilities. While it uses RDDs internally, it's not the direct abstraction for parallel execution.

2. What component resides on top of Spark Core?

A) Spark Streaming

B) Spark SQL

C) RDDs

D) None of the above

Ans - B) Spark SQL

Option A: Spark Streaming - Incorrect. Spark Streaming is a component that provides stream processing capabilities and builds on top of Spark Core, but it is not the answer to this specific question, as Spark SQL is more directly related to structured data processing.

Option B: Spark SQL - Correct. Spark SQL is a component that resides on top of Spark Core. It provides a higher-level API for querying structured data using SQL syntax and integrates with Spark's core functionalities through DataFrames and Datasets.

Option C: RDDs - Incorrect. RDDs are the fundamental abstraction in Spark Core for parallel execution and distributed processing. They are not a component that resides on top of Spark Core but rather part of the core abstraction.

Option D: None of the above - Incorrect. Spark SQL is indeed a component that resides on top of Spark Core, making this option incorrect.

3. Which statements about Cassandra and its Snitches are correct?

Statement 1: In Cassandra, during a write operation, when a hinted handoff is enabled and if any replica is down, the coordinator writes to all other replicas and keeps the write locally until the down replica comes back up.

Statement 2: In Cassandra, Ec2Snitch is an important snitch for deployments, and it is a simple snitch for Amazon EC2 deployments where all nodes are in a single region. In Ec2Snitch, the region name refers to the data center, and the availability zone refers to the rack in a cluster.

A) Only Statement 1 is correct.

B) Only Statement 2 is correct.

C) Both Statement 1 and Statement 2 are correct.

D) Neither Statement 1 nor Statement 2 is correct.

Ans- C) Both Statement 1 and Statement 2 are correct.

Statement 1: Correct. In Cassandra, when a hinted handoff is enabled, if any replica is down during a write operation, the coordinator writes to all other available replicas and keeps a hint for the down replica. Once the down replica comes back online, the coordinator will hand off the hinted write to that replica.

Statement 2: Correct. Ec2Snitch is a snitch used in Cassandra for Amazon EC2 deployments. It assumes that all nodes are within a single region. In Ec2Snitch, the term "region" corresponds to the data center, and "availability zone" corresponds to the rack within the cluster, helping to optimize data placement and replication.

4.Which of the following is a module for Structured data processing?

a. GraphX
b. MLlib
c. Spark SQL
d. Spark R

Option a: GraphX- Incorrect. This module is for graph processing and analytics, allowing for the manipulation and analysis of graph data structures.
Option b: MLlib- Incorrect. This is Spark’s machine learning library, providing algorithms and utilities for machine learning tasks, not specifically for structured data processing.
Option c: Spark SQL- Correct. Spark SQL is the module designed specifically for structured data processing. It provides a programming interface for working with structured and semi-structured data. It allows querying of data via SQL, integrates with DataFrames and Datasets, and provides optimizations through the Catalyst optimizer and Tungsten execution engine.

Option d: Spark R - Incorrect. This module provides support for using R with Spark, primarily aimed at statistical computing and data analysis rather than structured data processing specifically.

5. A healthcare provider wants to store and query patient records in a NoSQL database with high write throughput and low-latency access. Which Hadoop ecosystem technology is most suitable for this requirement?

A) Apache Hadoop

B) Apache Spark

C) Apache HBase

D) Apache Pig

Ans- C) Apache HBase

Option A: Apache Hadoop - Incorrect. Apache Hadoop is a framework for distributed storage and processing of large data sets using the Hadoop Distributed File System (HDFS) and MapReduce. It is not specifically optimized for low-latency access or high write throughput.

Option B: Apache Spark - Incorrect. Apache Spark is a fast, in-memory data processing engine that can handle large-scale data analytics and processing. While it offers low-latency data processing, it is not a NoSQL database and is not designed primarily for high write throughput.

Option C: Apache HBase - Correct. Apache HBase is a distributed, scalable, NoSQL database that runs on top of HDFS. It is designed for high write throughput and low-latency access to large volumes of data, making it suitable for storing and querying patient records efficiently.

Option D: Apache Pig - Incorrect. Apache Pig is a high-level platform for creating MapReduce programs used with Hadoop. It is primarily used for data transformation and analysis, not for high write throughput or low-latency NoSQL data storage.

6.The primary Machine Learning API for Spark is now the _____ based API

a. DataFrame
b. Dataset
c. RDD
d. All of the above

Option A: DataFrame - Correct. The primary Machine Learning API for Spark is now based on DataFrames. Spark's MLlib, the machine learning library, has adopted DataFrames as the primary API for building and training machine learning models. This approach provides a higher-level API and better integration with Spark SQL, offering optimized performance and ease of use.

Option B: Dataset - Incorrect. While Datasets are a powerful API in Spark that provides type safety and functional programming constructs, the primary Machine Learning API is not based on Datasets. Instead, DataFrames are used.

Option C: RDD - Incorrect. RDDs (Resilient Distributed Datasets) were the original abstraction in Spark and were used in earlier versions of MLlib. However, the primary Machine Learning API has shifted to DataFrames for better integration and performance.

Option D: All of the above - Incorrect. While RDDs, DataFrames, and Datasets are all important abstractions in Spark, the primary Machine Learning API is now specifically based on DataFrames

7. How does Apache Spark's performance compare to Hadoop MapReduce?

a) Apache Spark is up to 10 times faster in memory and up to 100 times faster on disk.

b) Apache Spark is up to 100 times faster in memory and up to 10 times faster on disk.

c) Apache Spark is up to 10 times faster both in memory and on disk compared to Hadoop MapReduce.

d) Apache Spark is up to 100 times faster both in memory and on disk compared to Hadoop MapReduce.

Ans- b) Apache Spark is up to 100 times faster in memory and up to 10 times faster on disk.

Option a: Incorrect. While Spark is indeed faster than Hadoop MapReduce, the comparison numbers are not accurate. Spark can be up to 100 times faster in memory, but the figure of 10 times faster on disk is not the correct characterization.

Option b: Correct. Apache Spark is known for its significant performance improvements over Hadoop MapReduce. It can be up to 100 times faster when processing data in memory, due to its in-memory computation capabilities. On disk, Spark is up to 10 times faster compared to MapReduce because of its efficient data processing and reduced disk I/O.

Option c: Incorrect. Spark is not just 10 times faster both in memory and on disk. It achieves up to 100 times faster performance in memory and up to 10 times faster on disk.

Option d: Incorrect. While Spark can be up to 100 times faster in memory, it is not typically characterized as being up to 100 times faster on disk. The correct comparison indicates up to 10 times faster on disk.

8.Which DAG action in Apache Spark triggers the execution of all previously defined transformations in the DAG and returns the count of elements in the resulting RDD or DataFrame?

a. collect()
b. count()
c. take()
d. first()

Option a: collect() - **Incorrect.** The `collect()` action triggers the execution of all previously defined transformations and retrieves all elements of the RDD or DataFrame to the driver program. It does not return the count of elements but rather returns the complete dataset.

Option b: count() - **Correct.** The `count()` action triggers the execution of all previously defined transformations in the DAG and returns the number of elements in the resulting RDD or DataFrame. It is specifically designed to return the count of elements.

Option c: take() - **Incorrect.** The `take()` action triggers execution and retrieves a specified number of elements from the RDD or DataFrame, but it does not return the count of all elements.

Option d: first() - **Incorrect.** The `first()` action triggers execution and retrieves the first element of the RDD or DataFrame, but it does not return the count of elements.

9. What is Apache Spark Streaming primarily used for?

a. Real-time processing of streaming data
b. Batch processing of static datasets
c. Machine learning model training
d. Graph processing

Option a: Real-time processing of streaming data - Correct. Spark Streaming is designed for processing continuous streams of data in real-time.
Option b: Batch processing of static datasets - Incorrect. Batch processing is better suited for static datasets.
Option c: Machine learning model training - Incorrect. While Spark can be used for machine learning, Spark Streaming is specifically for streaming data.
Option d: Graph processing - Incorrect. Graph processing is another area where Spark can be used, but Spark Streaming is focused on streaming data.

10.Which of the following represents the smallest unit of data processed by Apache Spark Streaming?

a. Batch
b. Window
c. Micro-batch
d. Record

Option a: Batch - Incorrect. A batch is a collection of micro-batches.
Option b: Window - Incorrect. A window is a time interval used for processing data.
Option c: Micro-batch - Correct. Micro-batches are the smallest unit of data processed in Spark Streaming.
Option d: Record - Incorrect. A record is a single unit of data within a micro-batch.

Week -4

1.Which of the following statements about Bloom filters is true?

a. Bloom filters guarantee no false negatives

b. Bloom filters use cryptographic hashing functions

c. Bloom filters may produce false positives but no false negatives

d. Bloom filters are primarily used for sorting large datasets

Option a: Bloom filters guarantee no false negatives - **Incorrect**. Bloom filters can produce false positives (indicating an element is present when it's not), but they guarantee no false negatives (indicating an element is absent when it's present).

Option b: Bloom filters use cryptographic hashing functions - **Incorrect**. While cryptographic hashing functions can be used, they are not a requirement. Bloom filters typically use multiple hash functions.

Option c: Bloom filters may produce false positives but no false negatives - **Correct**. This is the fundamental property of Bloom filters.

Option d: Bloom filters are primarily used for sorting large datasets - **Incorrect**. Bloom filters are primarily used for approximate membership testing.

2. How does CAP theorem impact the design of distributed systems?

A) It emphasizes data accuracy over system availability

B) It requires trade-offs between consistency, availability, and partition tolerance

C) It prioritizes system performance over data security

D) It eliminates the need for fault tolerance measures

Option A: It emphasizes data accuracy over system availability - Incorrect. The CAP theorem does not prioritize data accuracy; rather, it highlights the trade-offs between consistency, availability, and partition tolerance in distributed systems.

Option B: It requires trade-offs between consistency, availability, and partition tolerance - Correct. The CAP theorem states that in the presence of a network partition, a distributed system can only guarantee either consistency or availability, but not both.

Option C: It prioritizes system performance over data security - Incorrect. The CAP theorem does not address performance or security; it focuses specifically on the consistency, availability, and partition tolerance trade-offs in distributed systems.

Option D: It eliminates the need for fault tolerance measures - Incorrect. The CAP theorem does not eliminate the need for fault tolerance; in fact, it highlights the challenges that arise in maintaining consistency and availability when partitions occur.

3. Which guarantee does the CAP theorem consider as mandatory for a distributed system?

a. Consistency
b. Availability
c. Partition tolerance
d. Latency tolerance

Option a: Consistency - **Incorrect**. The CAP theorem states that it's impossible to achieve all three guarantees (Consistency, Availability, and Partition tolerance) simultaneously in a distributed system.

Option b: Availability - **Incorrect**. Availability is not guaranteed if partition tolerance is required.

Option c: Partition tolerance - **Correct**. The CAP theorem states that partition tolerance is essential for distributed systems, as network partitions are inevitable.

Option d: Latency tolerance - **Incorrect**. Latency tolerance is not explicitly mentioned in the CAP theorem.

4.What consistency level in Apache Cassandra ensures that a write operation is acknowledged only after the write has been successfully written to all replicas?

a. ONE
b. LOCAL_ONE
c. LOCAL_QUORUM
d. ALL

Option a: ONE - **Incorrect**. ONE requires only one replica to acknowledge the write.

Option b: LOCAL_ONE - **Incorrect**. LOCAL_ONE requires one replica within the same datacenter to acknowledge the write.

Option c: LOCAL_QUORUM - **Incorrect**. LOCAL_QUORUM requires a quorum of replicas within the same datacenter to acknowledge the write.

Option d: ALL - **Correct**. ALL requires all replicas to acknowledge the write before returning a response.

5. How does Zookeeper contribute to maintaining consistency in distributed systems?

A) By managing data replication

B) By providing a centralized configuration service

C) By ensuring data encryption

D) By optimizing data storage

Option A: By managing data replication – Incorrect. Zookeeper's role is more about coordination, not directly managing data replication.

Option B: By providing a centralized configuration service – Correct.

Zookeeper contributes to maintaining consistency in distributed systems by providing a centralized coordination and configuration service, ensuring consistent synchronization across distributed nodes.

Option C: By ensuring data encryption – **Incorrect.** Zookeeper doesn't handle encryption.

Option D: By optimizing data storage – **Incorrect.** Zookeeper is not involved in optimizing data storage.

Explanation:

Centralized Configuration: ZooKeeper acts as a centralized service where distributed applications can store and retrieve configuration information. This helps in ensuring that all nodes in the distributed system have consistent configuration settings, reducing the chances of configuration mismatches or inconsistencies.

6. A ___________ server is a machine that keeps a copy of the state of the entire system and persists this information in local log files.

a) Master
b) Region
c) Zookeeper
d) All of the mentioned

Option A: Master – **Incorrect.** The master server may manage parts of the system but does not persist the full system state in local logs.

Option B: Region – **Incorrect.** A region server manages a subset of data, but it doesn't maintain the full system state or persist it in logs.

Option C: Zookeeper – **Correct.** Zookeeper maintains a consistent view of the system's state and stores this information in local log files for fault tolerance and recovery.

Option D: All of the mentioned – **Incorrect.** Only Zookeeper is responsible for persisting the state of the entire system in local logs.

Explanation:

Master Server: A master server typically coordinates tasks within a cluster but doesn't necessarily store the entire system state.

Region Server: This term is often used in context of distributed databases like HBase, where region servers manage specific data partitions. They wouldn't hold the entire system state.

Zookeeper

Zookeeper is a centralized service that coordinates and manages distributed systems. It keeps a copy of the system's state and persists this information in local log files. This allows it to provide services such as naming, configuration management, and synchronization.

While a Master node might also have some state information, its primary role is often different, such as coordinating tasks or managing data. A Region node is typically a unit within a larger distributed system, and its role might involve managing specific data or tasks.

7.What is Apache Zookeeper primarily used for in Big Data ecosystems?

A) Data storage

B) Data processing

C) Configuration management

D) Data visualization

Option A: Data storage – **Incorrect.** Zookeeper is not designed for storing large amounts of data; its main purpose is coordination, not data storage.

Option B: Data processing – **Incorrect.** Zookeeper does not process data; it provides coordination services for distributed systems.

Option C: Configuration management – **Correct.** Zookeeper is primarily used for configuration management, leader election, and synchronization in distributed systems within Big Data ecosystems.

Option D: Data visualization – **Incorrect.** Zookeeper has no role in data visualization. Its function is more about system coordination and management.

8. Which statement correctly describes CQL (Cassandra Query Language)?

   a. CQL is a SQL-like language used for querying relational databases
   b. CQL is a procedural programming language used for writing stored procedures in Cassandra
   c. CQL is a language used for creating and managing tables and querying data in Apache Cassandra
   d. CQL is a scripting language used for data transformation tasks in Cassandra

   Option A: CQL is a SQL-like language used for querying relational databases – Incorrect. While CQL is SQL-like, Cassandra is a NoSQL database, not a relational database.

   Option B: CQL is a procedural programming language used for writing stored procedures in Cassandra – Incorrect. CQL is not a procedural language, nor is it used for writing stored procedures

.

   Option C: CQL is a language used for creating and managing tables and querying data in Apache Cassandra – Correct. CQL is primarily used in Cassandra for creating, managing tables, and querying data.

   Option D: CQL is a scripting language used for data transformation tasks in Cassandra – Incorrect. CQL is not a scripting language and is not designed for data transformation tasks.

9.Which aspect of CAP theorem refers to a system's ability to continue operating despite network failures?

A) Consistency

B) Accessibility

C) Partition tolerance

D) Atomicity

Option A: Consistency – Incorrect. Consistency refers to ensuring that all nodes see the same data at the same time, not handling network failures.

Option B: Accessibility – **Incorrect.** Availability refers to the system's ability to respond to requests, but does not specifically address network partitioning.

Option C: Partition tolerance – **Correct.** Partition tolerance refers to the system's ability to continue functioning even when network failures or partitions occur.

Option D: Atomicity – **Incorrect.** Atomicity is a concept related to transactions, ensuring that operations are fully completed or not at all, not related to network failures.

10. Why are tombstones used in distributed databases like Apache Cassandra?
    a. To mark nodes that are temporarily unavailable
    b. To mark data that is stored in multiple replicas
    c. To mark data that has been logically deleted
    d. To mark data that is actively being updated

    Option a: To mark nodes that are temporarily unavailable - **Incorrect**. Tombstones are not used to mark unavailable nodes.

    Option b: To mark data that is stored in multiple replicas - **Incorrect**. Tombstones are not used to mark data replication.

    Option c: To mark data that has been logically deleted - **Correct**. Tombstones are used to mark data that has been deleted but still exists in the system for a certain period to prevent accidental overwrites.

    Option d: To mark data that is actively being updated - **Incorrect**. Tombstones are not used to mark data that is being updated.

Week 5

1. What distributed graph processing framework operates on top of Spark?
    a. MLlib
    b. GraphX
    c. Spark streaming
    d. ALL

a. **Incorrect -MLlib**:

   This is Spark’s machine learning library, not specifically for graph processing.

b. **Correct- GraphX**

   It is a distributed graph processing framework that operates on top of Apache Spark. It allows users to process and analyze graph data at scale by leveraging Spark's core functionality for distributed data processing.

c. **Incorrect -Spark Streaming**:

   This is used for processing real-time data streams, not for graph processing.

d. **ALL**:

   Incorrect, as only GraphX is specifically designed for graph processing on Spark.

2.Which of the following frameworks is best suited for fast, in-memory data processing and supports advanced analytics such as machine learning and graph processing?

a) Apache Hadoop MapReduce
b) Apache Flink
c) Apache Storm
d) Apache Spark

a) Apache Hadoop MapReduce

- **Incorrect**: Apache Hadoop MapReduce is primarily designed for batch processing and relies on disk-based storage, making it slower compared to in-memory processing. It is not specifically designed for advanced analytics like machine learning or graph processing.

b) Apache Flink

- **Incorrect**: Apache Flink is excellent for real-time stream processing and supports complex event processing. However, it is primarily optimized for stream processing rather than providing a comprehensive suite for batch processing, machine learning, and graph analytics as Spark does.

c) Apache Storm

- **Incorrect**: Apache Storm focuses on real-time stream processing with low latency. While it excels in handling real-time data, it does not offer the same breadth of support for in-memory processing, machine learning, and graph processing as Spark.

d)**Apache Spark**

- **correct**:**Apache Spark** is designed for fast, in-memory data processing. It provides advanced analytics capabilities, including machine learning (through MLlib) and graph processing (through GraphX). Its in-memory computation greatly speeds up processing tasks compared to disk-based systems. Spark's flexibility and performance make it ideal for handling complex analytics and iterative algorithms efficiently.

3. A financial institution needs to analyze historical stock market data to predict market trends and make investment decisions. Which Big Data processing framework is best suited for this scenario?

a. Apache Spark
b. Apache Storm
c. Hadoop MapReduce
d. Apache Flume

a) Correct- Apache Spark

Explanation: Apache Spark is well-suited for analyzing historical data due to its fast in-memory processing capabilities. It supports a variety of data analytics tasks and is ideal for predictive analytics and machine learning.

b) **Incorrect- Apache Storm**: This is for real-time stream processing, which is less suited for historical data analysis.

c) **Incorrect- Hadoop MapReduce**: While it can handle large-scale data processing, it is slower than Spark for iterative algorithms used in predictive modeling.

d) **Incorrect- Apache Flume**: It is primarily used for data ingestion, not analysis.

4.A telecommunications company needs to process real-time call logs from millions of subscribers to detect network anomalies. Which combination of Big Data tools would be appropriate for this use case?

a. Apache Hadoop and Apache Pig
b. Apache Kafka and Apache HBase
c. Apache Spark and Apache Hive
d. Apache Storm and Apache pig

a. **Incorrect**- **Apache Hadoop and Apache Pig**: Hadoop is designed for batch processing, and Pig is used for ETL tasks in batch mode. This combination is not suited for real-time processing.

b. **Correct**- **Apache Kafka and Apache HBase**

- **Apache Kafka**: Kafka is a distributed event streaming platform that excels at handling real-time data streams. It can ingest high volumes of log data efficiently and serve as a buffer for processing.
- **Apache HBase**: HBase is a NoSQL database that provides real-time read/write access to large datasets. It complements Kafka by offering a scalable and distributed storage solution where processed data can be stored and queried quickly.

c. **Incorrect**- **Apache Spark and Apache Hive**: While Spark can handle real-time processing (with Structured Streaming) and is powerful for analytics, Hive is more oriented towards batch processing and querying rather than real-time analytics.

d. **Incorrect**- **Apache Storm and Apache Pig**: Storm is good for real-time stream processing, but Pig is used for batch processing and ETL tasks. Combining Storm with Pig does not align with the need for real-time analytics and storage.

5. Do many people use Kafka as a substitute for which type of solution?

a. log aggregation
b. Compaction

c. Collection
d. all of the mentioned

a. **Correct**- **Log aggregation-** Apache Kafka is commonly used for log aggregation. It efficiently collects, processes, and stores log data from various sources in a distributed manner.

b. **Incorrect -Compaction**: Kafka supports log compaction, but it's not a substitute for log aggregation.

c. **Incorrect -Collection**: Kafka is used for collecting logs, but the term 'substitute' more directly refers to log aggregation.

d. **Incorrect -All of the mentioned**: Not all options are directly applicable.

6. Which of the following features of Resilient Distributed Datasets (RDDs) in Apache Spark contributes to their fault tolerance?

a. DAG (Directed Acyclic Graph)
b. In-memory computation
c. Lazy evaluation
d. Lineage information

a. **Incorrect - DAG (Directed Acyclic Graph)**: While the DAG is essential for understanding how Spark schedules tasks, it is the lineage information specifically that ensures fault tolerance.

b. **Incorrect - In-memory computation**: This improves performance but does not directly contribute to fault tolerance.

c. **Incorrect -Lazy evaluation**: This optimizes execution and resource usage but does not specifically address fault tolerance.

d. **correct -Lineage information**: RDDs maintain lineage information, which is the history of transformations applied to the data. This lineage information allows Spark to recompute lost data in case of failures, ensuring fault tolerance.

7.Point out the correct statement.

a. Hadoop do need specialized hardware to process the data
b. Hadoop allows live stream processing of real-time data

c. In the Hadoop mapreduce programming framework output files are divided into lines or records
d. None of the mentioned

a. Hadoop do need specialized hardware to process the data

- **Incorrect**: Hadoop is designed to run on commodity hardware, meaning it does not require specialized or high-end hardware. It is intended to scale across many inexpensive, standard machines.

b. Hadoop allows live stream processing of real-time data

- **Incorrect**: Traditional Hadoop, specifically the MapReduce framework, is designed for batch processing and does not natively support real-time stream processing. For real-time processing, frameworks like Apache Storm or Apache Flink are more appropriate.

c. In the Hadoop MapReduce programming framework output files are divided into lines or records

- **Correct**: In Hadoop MapReduce, the output is indeed processed and written as lines or records, where each record represents a unit of data processed by the framework. This is how data is commonly handled in MapReduce jobs.

d. None of the mentioned

- **Incorrect**: The third statement is accurate regarding how Hadoop MapReduce processes and outputs data.

8. Which of the following statements about Apache Pig is true?

a. Pig Latin scripts are compiled into HiveQL for execution.
b. Pig is primarily used for real-time stream processing.
c. Pig Latin provides a procedural data flow language for ETL tasks.
d. Pig uses a schema-on-write approach for data storage.

a. **Incorrect**: **Pig Latin scripts are compiled into HiveQL for execution**:

- Pig Latin is compiled into a series of MapReduce jobs.

b. **Incorrect**: **Pig is primarily used for real-time stream processing**:

- Pig is used for batch processing.

c. **correct - Pig Latin provides a procedural data flow language for ETL tasks.** Pig Latin is a scripting language used in Apache Pig for expressing data transformation tasks in a procedural manner, making it suitable for ETL (Extract, Transform, Load) processes.

d. **Incorrect**: **Pig uses a schema-on-write approach for data storage**:

- Pig uses schema-on-read.

9. An educational institution wants to analyze student performance data stored in HDFS and generate personalized learning recommendations. Which Hadoop ecosystem components should be used?

a. Apache HBase for storing student data and Apache Pig for processing.
b. Apache Kafka for data streaming and Apache Storm for real-time analytics.
c. Hadoop MapReduce for batch processing and Apache Hive for querying.
d. Apache Spark for data processing and Apache Hadoop for storage.

a. **Incorrect - Apache HBase for storing student data and Apache Pig for processing**: While HBase is a NoSQL database suitable for real-time read/write access, Pig is used for ETL tasks and batch processing, which may not be as efficient as Spark for complex analytics and recommendations.

b. **Incorrect - Apache Kafka for data streaming and Apache Storm for real-time analytics**: Kafka is used for streaming data, and Storm is for real-time analytics. This combination is more suited for real-time data processing rather than batch analytics and recommendation generation.

c. **Incorrect - Hadoop MapReduce for batch processing and Apache Hive for querying**: While MapReduce and Hive are both part of the Hadoop ecosystem, MapReduce is less efficient for iterative processing compared to Spark. Hive is used for querying but is more oriented towards batch processing rather than real-time analytics and personalized recommendations.

d. Correct - Apache Spark for data processing and Apache Hadoop for storage.

- **Apache Spark for data processing**: Spark is a powerful and versatile data processing engine that supports complex analytics, machine learning, and iterative algorithms. It is well-suited for analyzing large datasets and generating recommendations. Spark's in-memory computation capabilities provide high performance for such tasks.

- **Apache Hadoop for storage**: Hadoop's HDFS (Hadoop Distributed File System) is a scalable and reliable storage system designed for storing large volumes of data across a distributed cluster. It is ideal for storing the large datasets of student performance data.

10.A company is analyzing customer behavior across multiple channels (web, mobile app, social media) to personalize marketing campaigns. Which technology is best suited to handle this type of data processing?

a. Hadoop MapReduce
b. Apache Kafka
c. Apache Spark
d. Apache Hive

a. **Incorrect - Hadoop MapReduce**: While it can handle large-scale data, it is less efficient for iterative and real-time analytics compared to Spark.

b. **Incorrect - Apache Kafka**: Primarily used for message streaming, not data processing.

c. correct - Apache Spark - Apache Spark is highly suitable for analyzing customer behavior across various channels due to its fast processing capabilities and support for complex analytics and machine learning.

d. **Incorrect - Apache Hive:** Used for querying and not for complex data processing and analytics.

Week 6

1. Point out the wrong statement.

   a) Replication Factor can be configured at a cluster level (Default is set to 3) and also at a file level

   b) Block Report from each DataNode contains a list of all the blocks that are stored on that DataNode
   c) User data is distributed across multiple DataNodes in the cluster and is managed by the NameNode.
   d) DataNode is aware of the files to which the blocks stored on it belong to

a) Replication Factor can be configured at a cluster level (Default is set to 3) and also at a file level

- **Correct**. Replication Factor can indeed be set at both the cluster level and the file level in distributed file systems like HDFS.

b) Block Report from each DataNode contains a list of all the blocks that are stored on that DataNode

- **Correct**. A Block Report includes a list of all blocks stored on a DataNode.

c) User data is distributed across multiple DataNodes in the cluster and is managed by the NameNode.

- **correct**. In a distributed file system like HDFS, user data is stored in a distributed manner across the cluster and managed by the HDFS, not just the local file system of each DataNode.

d) DataNode is aware of the files to which the blocks stored on it belong to

- **Incorrect**. DataNodes manage blocks of data and are not aware of the higher-level file structure; this information is managed by the NameNode in HDFS.

2. What is the primary technique used by Random Forest to reduce overfitting?

   a) Boosting

   b) Bagging

   c) Pruning

   d) Neural networks

a. **Incorrect - Boosting:**

- **Not Used in Random Forest:** Boosting is a different technique used in methods like Gradient Boosting, where trees are built sequentially to correct errors from previous trees. It's not used by Random Forest, which relies on bagging.

b. **Correct - Bagging (Bootstrap Aggregating):**

- **Primary Technique Used by Random Forest:** Bagging is the key technique used by Random Forest to reduce overfitting. In Random Forest, multiple decision trees are trained on different random subsets of the data, created through bootstrapping (sampling with replacement). Each tree is trained independently on its subset of data, and the predictions from all trees are aggregated (typically by voting or averaging) to produce the final result. This approach helps to reduce variance and improves the model's ability to generalize by averaging out the errors from individual trees.

c. **Incorrect - Pruning:**

- **Not a Primary Technique in Random Forest:** Pruning is a technique used to reduce the size of decision trees by removing parts that are not contributing to the prediction accuracy. While pruning helps to control overfitting in individual decision trees, Random Forest primarily relies on bagging for overfitting reduction.

d. **Incorrect - Neural Networks:**

- **Not Related to Random Forest:** Neural networks are a different class of models and are not related to the ensemble method of Random Forest.

3. What statements accurately describe Random Forest and Gradient Boosting ensemble methods?

   S1: Both methods can be used for classification task

   S2: Random Forest is use for regression whereas Gradient Boosting is use for Classification task

   S3: Random Forest is use for classification whereas Gradient Boosting is use for regression task

   S4: Both methods can be used for regression

A) S1 and S2
B) S2 and S4
C) S3 and S4
D) S1 and S4

S1: Both methods can be used for classification tasks

- **Correct**. Both Random Forest and Gradient Boosting can be used for classification problems.

S2: Random Forest is used for regression whereas Gradient Boosting is used for Classification task

- **Incorrect**. Random Forest and Gradient Boosting can both be used for both regression and classification tasks.

S3: Random Forest is used for classification whereas Gradient Boosting is used for regression task

- **Incorrect**. As with S2, both methods can be used for both types of tasks.

S4: Both methods can be used for regression

- **Correct**. Both Random Forest and Gradient Boosting can be used for regression tasks as well.

4. In the context of K-means clustering with MapReduce, what role does the **Map** phase play in handling very large datasets?

A) It reduces the size of the dataset by removing duplicates

B) It distributes the computation of distances between data points and centroids across multiple nodes

C) It initializes multiple sets of centroids to improve clustering accuracy

D) It performs principal component analysis (PCA) on the data

A) It reduces the size of the dataset by removing duplicates

- **Incorrect**. The Map phase does not focus on removing duplicates but rather on distributing and processing the data.

B) It distributes the computation of distances between data points and centroids across multiple nodes

- **Correct**. The Map phase is responsible for calculating distances between data points and centroids and distributing this task across nodes.

C) It initializes multiple sets of centroids to improve clustering accuracy

- **Incorrect**. Initialization of centroids is generally done before the Map phase starts and is not part of its functionality.

D) It performs principal component analysis (PCA) on the data

- **Incorrect**. PCA is not typically done in the Map phase; it is a preprocessing step for dimensionality reduction.

5. What is a common method to improve the performance of the K-means algorithm when dealing with large-scale datasets in a MapReduce environment?

A) Using hierarchical clustering before K-means

B) Reducing the number of clusters

C) Employing mini-batch K-means

D) Increasing the number of centroids

A) Using hierarchical clustering before K-means

- **Incorrect**. While hierarchical clustering can be used to initialize centroids, it is not specific to improving K-means in a MapReduce environment.

B) Reducing the number of clusters

- **Incorrect**. Reducing the number of clusters might not improve performance and could lead to less meaningful clustering.

C) Employing mini-batch K-means

- **Correct**. Mini-batch K-means is a method used to handle large-scale datasets efficiently by processing small, random subsets of the data.

D) Increasing the number of centroids

- **Incorrect**. Increasing the number of centroids might not improve performance and could complicate the clustering process.

6.Which similarity measure is often used to determine the similarity between two text documents by considering the angle between their vector representations in a high-dimensional space?

A) Manhattan Distance

B) Cosine Similarity

C) Jaccard Similarity

D) Hamming Distance

A) Manhattan Distance

- **Incorrect**. Manhattan Distance is not used for text document similarity in this context.

B) Cosine Similarity

- **Correct**. Cosine Similarity measures the cosine of the angle between two vectors, making it ideal for text documents in high-dimensional space.

C) Jaccard Similarity

- **Incorrect**. Jaccard Similarity is used for comparing sets and is not based on vector angles.

D) Hamming Distance

- **Incorrect**. Hamming Distance is used for comparing strings of equal length and is not applicable to text document similarity in vector space.

7.Which distance measure calculates the distance along strictly horizontal and vertical paths, consisting of segments along the axes?

A)Minkowski distance

B) Cosine similarity

c) Manhattan distance

D) Euclidean distance

A) Minkowski distance

- **Incorrect**. Minkowski distance generalizes Euclidean and Manhattan distances but is not specific to axis-aligned paths.

B) Cosine similarity

- **Incorrect**. Cosine similarity measures the angle between vectors and does not involve distance calculation.

C) Manhattan distance

- **Correct**. Manhattan distance measures distance along axis-aligned paths (horizontal and vertical segments).

D) Euclidean distance

- **Incorrect**. Euclidean distance measures the straight-line distance between points, not restricted to axis-aligned paths.

8.What is the purpose of a **validation set** in machine learning?

A) To train the model on unseen data

B) To evaluate the model's performance on the training data

C) To tune hyperparameters and prevent overfitting

D) To test the final model's performance

A) To train the model on unseen data

- **Incorrect**. The validation set is not used for training but for model evaluation during training.

B) To evaluate the model's performance on the training data

- **Incorrect**. Evaluation on training data is not the purpose of a validation set; it is used for hyperparameter tuning and model selection.

C) To tune hyperparameters and prevent overfitting

- **Correct**. The validation set is used to tune hyperparameters and monitor performance to avoid overfitting.

D) To test the final model's performance

- **Incorrect**. Testing the final model's performance is done using a separate test set, not the validation set.

9. In **K-fold cross-validation**, what is the purpose of splitting the dataset into K folds?

A) To ensure that every data point is used for training only once

B) To train the model on all the data points

C) To test the model on the same data multiple times

D) To evaluate the model’s performance on different subsets of data

A) To ensure that every data point is used for training only once

- **Incorrect**. In K-fold cross-validation, every data point is used for training multiple times, not just once.

B) To train the model on all the data points

- **Incorrect**. The dataset is split into folds, and only K-1 folds are used for training each time.

C) To test the model on the same data multiple times

- **Incorrect**. Each fold is used for testing once, and training is done on the remaining folds.

D) To evaluate the model’s performance on different subsets of data

- **Correct**. K-fold cross-validation ensures that the model is evaluated on different subsets of the data, providing a robust measure of its performance.

10. Which of the following steps is NOT typically part of the machine learning process?

A) Data Collection

B) Model Training

C) Model Deployment

D) Data Encryption

A) Data Collection

- **Incorrect**. Data Collection is a fundamental step in the machine learning process.

B) Model Training

- **Incorrect**. Model Training is a core step in machine learning.

C) Model Deployment

- **Incorrect**. Model Deployment is part of the machine learning lifecycle, as it involves putting the model into production.

D) Data Encryption

- **Correct**. Data Encryption is not typically a part of the machine learning process itself, though it may be relevant for data security and privacy.

Week 7

1. **What is the primary purpose of using a decision tree in regression tasks within big data environments?**

A) To classify data into distinct categories
B) To predict continuous values based on input features
C) To reduce the dimensionality of the dataset
D) To perform clustering of similar data points

Answer:

A) To classify data into distinct categories:

- **Incorrect:** This is the primary purpose of decision trees in classification tasks, not regression. Classification involves predicting discrete labels or categories rather than continuous values.

B) To predict continuous values based on input features:

- **Correct- Primary Purpose in Regression Tasks:** In regression tasks, a decision tree is used to predict continuous values (e.g., predicting house prices, stock prices, etc.) based on input features. The decision tree splits the data into different branches based on feature values, aiming to minimize the variance within each branch to make accurate predictions for continuous outcomes.

C) To reduce the dimensionality of the dataset:

- **Incorrect:** Dimensionality reduction is typically performed by techniques like Principal Component Analysis (PCA), not decision trees. Decision trees do not inherently reduce the number of features but rather use them to make predictions.

D) To perform clustering of similar data points:

- **Incorrect:** Clustering is a technique used to group similar data points together and is typically performed by algorithms such as K-means or hierarchical clustering. Decision trees are not used for clustering tasks.

2.Which statement accurately explains the function of bootstrapping within the random forest algorithm?

A) Bootstrapping creates additional features to augment the dataset for improved random forest performance.
B) Bootstrapping is not used in the random forest algorithm, it is only employed in decision tree construction.
C) Bootstrapping produces replicas of the dataset by random sampling with replacement, which is essential for the random forest algorithm.
D) Bootstrapping generates replicas of the dataset without replacement, ensuring diversity in the random forest.

Answer:

A) **Incorrect - Bootstrapping creates additional features:** This is not the purpose of bootstrapping.
B) **Incorrect - Bootstrapping is not used in random forest:** Bootstrapping is a fundamental component of the random forest algorithm.
C) **correct -** Bootstrapping produces replicas of the dataset by random sampling with replacement, which is essential for the random forest algorithm.

   Explain:In the Random Forest algorithm, bootstrapping involves creating multiple subsets of the original dataset by randomly sampling with replacement. Each decision tree in the Random Forest is trained on a different bootstrapped subset of the data. This technique helps to introduce variability among the individual trees, which contributes to the overall robustness and generalization of the Random Forest model.

D) **Incorrect - Bootstrapping generates replicas without replacement:** This would not introduce diversity and might lead to biased models.

3. In a big data scenario using MapReduce, how is the decision tree model typically built?

A) By using a single-node system to fit the model
B) By distributing the data and computations across multiple nodes for parallel processing
C) By manually sorting data before applying decision tree algorithms
D) By using in-memory processing on a single machine

Answer:

A) By using a single-node system to fit the model:

- **Incorrect:** Using a single-node system is not suitable for big data scenarios because it does not scale well with large datasets. MapReduce is specifically designed to work with distributed systems to handle large-scale data processing.

B) By distributing the data and computations across multiple nodes for parallel processing:

- **Correct:** In a big data scenario using MapReduce, building a decision tree model involves distributing the data and computations across multiple nodes to leverage parallel processing. This approach allows for efficient handling of large datasets by dividing the work among several nodes in a cluster. Each node processes a portion of the data, and the results are aggregated to construct the final decision tree model.

C) By manually sorting data before applying decision tree algorithms:

- **Incorrect:** Manual sorting of data is not a typical requirement for decision tree algorithms in a MapReduce framework. MapReduce handles data partitioning and sorting automatically during the Map and Reduce phases.

D) By using in-memory processing on a single machine:

- **Incorrect:** In-memory processing on a single machine is not practical for big data scenarios due to memory limitations and scalability issues. MapReduce processes data distributed across multiple nodes, which is essential for handling large datasets effectively.

4. In Apache Spark, what is the primary purpose of using cross-validation in machine learning pipelines?

A) To reduce the number of features used in the model
B) To evaluate the model's performance by partitioning the data into training and validation sets multiple times
C) To speed up the data preprocessing phase
D) To increase the size of the training dataset by generating synthetic samples

Answer:

A) To reduce the number of features used in the model:

- **Incorrect:** Reducing the number of features is related to feature selection or dimensionality reduction techniques, such as Principal Component Analysis (PCA) or feature importance measures, rather than model evaluation.

B) Correct - To evaluate the model's performance by partitioning the data into training and validation sets multiple times:

- **Explanation:** This describes the process of **cross-validation**. Cross-validation is a technique used to evaluate a model's performance by partitioning the dataset into multiple subsets (or folds). The model is trained on some of these subsets and validated on the remaining ones. This process is repeated multiple times, each time with different subsets as the training and validation sets. This approach helps in assessing the model's performance more robustly and in mitigating issues related to overfitting or variability in the performance due to a single train-test split.

C) To speed up the data preprocessing phase:

- **Incorrect:** Techniques to speed up data preprocessing include data sampling, efficient algorithms, or parallel processing, but these are not directly related to model evaluation.

D) To increase the size of the training dataset by generating synthetic samples:

- **Incorrect:** This describes **data augmentation** or **synthetic data generation**, such as using techniques like SMOTE (Synthetic Minority Over-sampling

Technique) to balance datasets. It is not specifically about evaluating model performance through multiple partitions.

5. How does gradient boosting in machine learning conceptually resemble gradient descent in optimization theory?

A) Both techniques use large step sizes to quickly converge to a minimum
B) Both methods involve iteratively adjusting model parameters based on the gradient to minimize a loss function
C) Both methods rely on random sampling to update the model
D) Both techniques use a fixed learning rate to ensure convergence without overfitting

Answer:

A) Both techniques use large step sizes to quickly converge to a minimum:

- **Incorrect:** Gradient boosting and gradient descent do not necessarily use large step sizes. In fact, gradient boosting typically uses a smaller learning rate (step size) to ensure that the model converges slowly and avoids overfitting. Gradient descent can also use various step sizes, which are often tuned to balance convergence speed and stability.

B) Both methods involve iteratively adjusting model parameters based on the gradient to minimize a loss function:

Correct: Gradient boosting and gradient descent both use the concept of gradients to iteratively improve a model. In gradient boosting, new trees are added to the ensemble in a way that corrects the residual errors of the existing model, effectively adjusting the model to minimize the loss function. Each new tree is built to fit the negative gradient of the loss function, which is akin to taking steps in the direction of the steepest descent to reduce the overall error. Similarly, in gradient descent, the algorithm iteratively adjusts the parameters of a model by moving in the direction of the gradient of the loss function to find the minimum value.

C) Both methods rely on random sampling to update the model:

- **Incorrect:** Gradient boosting does not inherently rely on random sampling for updating the model; it builds trees sequentially to correct residuals. Gradient descent may involve stochastic or mini-batch sampling in its variants (such as stochastic gradient descent or mini-batch gradient descent), but this is not a direct conceptual similarity to gradient boosting.

D) Both techniques use a fixed learning rate to ensure convergence without overfitting:

- **Incorrect:** While gradient descent may use a fixed learning rate, gradient boosting typically uses a smaller learning rate as a regularization strategy to ensure gradual convergence and reduce the risk of overfitting. The learning rate is not fixed in the same sense for both methods; it is adjusted based on the context and specific implementation details.

6. Which statement accurately describes one of the benefits of decision trees?

A) Decision trees always outperform other models in predictive accuracy, regardless of the complexity of the dataset.
B) Decision trees can automatically handle feature interactions by combining different features within a single tree, but a single tree's predictive power is often limited.
C) Decision trees cannot handle large datasets and are not computationally scalable.
D) Decision trees require a fixed set of features and cannot adapt to new feature interactions during training.

Answer:

A) Decision trees always outperform other models in predictive accuracy, regardless of the complexity of the dataset:

- **Incorrect:** Decision trees do not always outperform other models. Their performance can vary based on the complexity of the dataset and other factors. They are prone to overfitting, particularly with complex datasets, and may not always provide the best predictive accuracy compared to other models like ensemble methods or neural networks.

B) Decision trees can automatically handle feature interactions by combining different features within a single tree, but a single tree's predictive power is often limited:

- **Correct:** Decision trees can indeed handle feature interactions automatically by splitting on different features at each node. This allows them to model complex relationships between features within the dataset. However, a single decision tree might not always have the predictive power or generalization capability, especially on its own, which is why ensemble methods like Random Forests or Gradient Boosting are often used to improve performance.

C) Decision trees cannot handle large datasets and are not computationally scalable:

- **Incorrect:** Decision trees can handle large datasets, but the computational resources and time required may increase with the size of the dataset. While they can be computationally intensive for very large datasets, they are scalable, and there are techniques and implementations designed to handle large-scale data.

D) Decision trees require a fixed set of features and cannot adapt to new feature interactions during training:

- **Incorrect:** Decision trees do not require a fixed set of features. They can adapt to new feature interactions dynamically as they grow. The structure of a decision tree is built by evaluating all available features to find the best splits at each node.

7. What has driven the development of specialized graph computation engines capable of inferring complex recursive properties of graph structured data?

A) Increasing demand for social media analytics
B) Advances in machine learning algorithms
C) Growing scale and importance of graph data
D) Expansion of blockchain technology

Answer:

Social Media Analytics (A): Incorrect: Social media analysis is a significant driver for graph technology adoption, but it's a specific application area benefiting from the capabilities of graph engines.

Machine Learning (B): Incorrect: Machine learning algorithms can benefit from graph data and graph computations, but the development of graph engines is not solely driven by advancements in machine learning.

(C) Graph-Structured Data: Correct: Many real-world relationships can be naturally modeled as graphs, where nodes represent entities (e.g., people, products) and edges represent connections between them (e.g., friendships, purchases).

Growing Scale: The amount of graph data is exploding due to social networks, recommendation systems, sensor networks, and other applications. Traditional relational databases struggle to handle the complexity and interconnectedness of graph data.

Blockchain (D): Incorrect: Blockchain technology utilizes some graph-like structures, but it's not the primary driver for the development of general-purpose graph computation engines.

8. Which of these statements accurately describes bagging in the context of understanding the random forest algorithm?

a) Bagging is primarily used to average predictions of decision trees in the random forest algorithm.

b) Bagging is a technique exclusively designed for reducing the bias in predictions made by decision trees.

c) Bagging, short for Bootstrap Aggregation, is a general method for averaging predictions of various algorithms, not limited to decision trees, and it works by reducing the variance of predictions.

d) Bagging is a method specifically tailored for improving the interpretability of decision trees in the random forest algorithm.

Answer:

a) **Incorrect - Bagging is primarily used to average predictions of decision trees in the random forest algorithm.**

- This is partially true but not entirely accurate. Bagging is used to reduce variance by averaging predictions, but it is a more general technique that can be applied to various models, not just decision trees.

b) **Incorrect - Bagging is a technique exclusively designed for reducing the bias in predictions made by decision trees.**

- This is incorrect. Bagging primarily aims to reduce variance rather than bias. In fact, while bagging can indirectly help reduce bias in some cases, its main benefit is in reducing the model's variance.

c) **Correct** - **Bagging, short for Bootstrap Aggregation, is a general method for averaging predictions of various algorithms, not limited to decision trees, and it works by reducing the variance of predictions.**

Explanation:

- **Bagging (Bootstrap Aggregation)** is indeed a technique designed to improve the stability and accuracy of machine learning algorithms by reducing variance. It involves generating multiple subsets of the data (through bootstrapping) and then training a separate model on each subset. The final prediction is typically an average (for regression) or a majority vote (for classification) of the predictions from each model. This method can be applied to various algorithms, not just decision trees, although it is particularly effective with high-variance models like decision trees.

d) **Incorrect - Bagging is a method specifically tailored for improving the interpretability of decision trees in the random forest algorithm.**

- This is incorrect. Bagging is not focused on improving interpretability but rather on improving the overall performance and stability of the model by reducing variance.

9.What is a key advantage of using regression trees in a big data environment when combined with MapReduce?

A) They require less computational power compared to other algorithms
B) They can handle both classification and regression tasks effectively
C) They automatically handle large-scale datasets by leveraging distributed processing
D) They eliminate the need for data preprocessing

Answer:

A) **Incorrect - They require less computational power compared to other algorithms:** While decision trees can be computationally efficient, they still require significant computational power for large-scale datasets.

B) **Incorrect - They can handle both classification and regression tasks effectively:** This is true, but it's not the primary advantage of using decision trees with MapReduce in a big data environment.

C) **Correct - They automatically handle large-scale datasets by leveraging distributed processing** is the key advantage of using regression trees in a big data environment when combined with MapReduce.

D) **Incorrect - They eliminate the need for data preprocessing:** Decision trees still require data preprocessing, such as handling missing values and feature scaling.

10. When implementing a regression decision tree using MapReduce, which technique helps in managing the data that needs to be split across different nodes?

A) Feature scaling
B) Data shuffling
C) Data partitioning
D) Model pruning

Answer:

A) **Incorrect - Feature scaling:** Feature scaling is used to normalize numerical features to a common range, which is important for many machine learning algorithms but not directly related to data partitioning in MapReduce.

B) **Incorrect - Data shuffling:** Data shuffling is the process of redistributing intermediate data between Map and Reduce tasks. While it's important for MapReduce jobs, it's not specific to decision trees.

C) **Correct - Data partitioning:** is the technique used to manage the data that needs to be split across different nodes in a MapReduce implementation of a regression decision tree.

D) **Incorrect - Model pruning:** Model pruning is a technique used to simplify a decision tree by removing unnecessary branches, which can improve its generalization performance. It's not directly related to data partitioning within MapReduce.

Week 8

1. **Which of the following statements accurately describes the functionality of a Parameter Server in the context of distributed machine learning?**

A) The Parameter Server handles data preprocessing by scaling features and normalizing values before training.

B) The Parameter Server distributes a model over multiple machines and provides two main operations: Pull (to query parts of the model) and Push (to update parts of the model).

C) The Parameter Server exclusively supports model training using (Stochastic) gradient descent and does not handle other machine learning algorithms.

D) The Parameter Server uses the Collapsed Gibbs Sampling method to update model parameters by aggregating push updates via subtraction.

Answer:

A) Incorrect - The Parameter Server handles data preprocessing by scaling features and normalizing values before training.
Explanation: Data preprocessing is generally performed by data processing pipelines or workers before the model is distributed to the Parameter Server, not by the Parameter Server itself.

B) Correct - The Parameter Server distributes a model over multiple machines and provides two main operations: Pull (to query parts of the model) and Push (to update parts of the model).
Explanation: A Parameter Server is a distributed system architecture that manages model parameters in a machine learning context. It enables multiple worker nodes to access and update the model parameters efficiently. The Pull operation allows workers to fetch the latest parameters, while the Push operation lets them send back the updates, such as gradients, to the server for model synchronization.

C) Incorrect - The Parameter Server exclusively supports model training using (Stochastic) gradient descent and does not handle other machine learning algorithms.

Explanation: Parameter Servers are versatile and can support various machine learning algorithms beyond gradient descent.

D) Incorrect - The Parameter Server uses the Collapsed Gibbs Sampling method to update model parameters.
Explanation: Parameter Servers typically rely on gradient-based optimization methods, not specifically on Gibbs Sampling.

2. **Why is PageRank considered important in the context of information retrieval on the World Wide Web?**

A) It helps to categorize web pages based on the quality of their content, thus improving the accuracy of search results.

B) It provides an objective and mechanical method for rating the importance of web pages based on the link structure of the web, addressing challenges of page relevance amidst a large number of web pages.

C) It ensures that all web pages are indexed equally, regardless of their content or link structure.

D) It automatically filters out irrelevant web pages by analyzing their content and metadata.

Answer:

A) Incorrect - It helps to categorize web pages based on the quality of their content.
Explanation: PageRank does not categorize pages based on content quality; instead, it focuses on their link structure.

B) Correct - It provides an objective and mechanical method for rating the importance of web pages based on the link structure of the web, addressing challenges of page relevance amidst a large number of web pages.
Explanation: PageRank assesses the importance of web pages by analyzing the structure of links between them. It ranks pages based on the quantity and quality of links they receive, helping search engines deliver relevant results in a vast and complex web landscape.

C) Incorrect - It ensures that all web pages are indexed equally, regardless of their content or link structure.
Explanation: PageRank prioritizes pages linked by other significant pages, meaning not all pages are treated equally.

D) Incorrect - It automatically filters out irrelevant web pages by analyzing their content and metadata.
Explanation: While content and metadata are considered in some search algorithms, PageRank primarily operates on link structure.

3.What role does the outerJoinVertices() operator serve in Apache Spark's GraphX?

A) It removes all vertices that are not present in the input RDD.

B) It returns a new graph with only the vertices from the input RDD.

C) It joins the input RDD data with vertices and includes all vertices, whether present in the input RDD or not.

D) It creates a subgraph from the input RDD and vertices.

Answer:

A) Incorrect - It removes all vertices that are not present in the input RDD.
Explanation: This operator does not exclude vertices; it retains them regardless of whether they have matching data.

B) Incorrect - It does not return a new graph with only the vertices from the input RDD.
Explanation: It includes all vertices from the original graph.

C) Correct - It joins the input RDD data with vertices and includes all vertices, whether present in the input RDD or not.
Explanation: The outerJoinVertices() operator allows for a join between vertex properties and RDD data, including all vertices in the graph. This is useful when you want to retain vertices that may not have corresponding data in the input RDD.

D) Incorrect - It does not create a subgraph from the input RDD and vertices.
Explanation: It creates a new graph that retains all vertices from the original graph and merges them with input data.

4. **Which of the following statements accurately describes a key feature of GraphX, a component built on top of Apache Spark Core?**

A) GraphX focuses exclusively on performing machine learning tasks and does not support graph processing.

B) GraphX allows for efficient graph processing and analysis, supports high-level graph measures like triangle counting, and integrates the Pregel API for graph traversal.

C) GraphX is primarily used for data ingestion and preprocessing and does not provide functionalities for graph algorithms or analytics.

D) GraphX provides only basic graph visualization capabilities and does not include algorithms like PageRank or triangle counting.

Answer:

A) Incorrect - GraphX focuses exclusively on performing machine learning tasks and does not support graph processing.
Explanation: GraphX is primarily focused on graph processing and not solely on machine learning tasks.

B) Correct - GraphX allows for efficient graph processing and analysis, supports high-level graph measures like triangle counting, and integrates the Pregel API for graph traversal.
Explanation: GraphX is designed for processing graphs at scale and includes functionalities for both analytical and algorithmic operations, such as triangle counting and custom graph traversals using the Pregel API.

C) Incorrect - GraphX is primarily used for data ingestion and preprocessing and does not provide functionalities for graph algorithms or analytics.
Explanation: While it can handle data ingestion, its main purpose is graph processing and analysis.

D) Incorrect - GraphX provides only basic graph visualization capabilities and does not include algorithms like PageRank or triangle counting.
Explanation: GraphX includes a variety of advanced graph algorithms, not just basic visualization capabilities.

5. Why are substantial indexes and data reuse important in graph processing?

A) To create decorative elements within graphs.

B) To save memory and processing resources by reusing routing tables and edge adjacency information.

C) To add redundancy to graphs for fault tolerance.

D) To increase the file size of graphs for better storage.

Answer:

A) Incorrect - To create decorative elements within graphs.
Explanation: Indexes and data reuse are focused on computational efficiency, not aesthetics.

B) Correct - To save memory and processing resources by reusing routing tables and edge adjacency information.
Explanation: Efficient graph processing relies on reducing redundancy and optimizing memory usage. Substantial indexes allow for quicker access to data without requiring excessive memory or processing power, leading to more efficient computations.

C) Incorrect - To add redundancy to graphs for fault tolerance.
Explanation: While redundancy can provide fault tolerance, substantial indexes are more about optimizing resource usage.

D) Incorrect - To increase the file size of graphs for better storage.
Explanation: Increasing file size is not a goal of using indexes or data reuse; rather, the aim is to reduce unnecessary data duplication.

6. Which of the following statement(s) accurately describe the functionality of operators in Apache Spark's GraphX?

A) Join operators add data to graphs and produce new graphs.

B) Structural operators operate on the structure of an input graph and produce a new graph.

C) Property operators modify the vertex or edge properties using a user-defined map function and produce a new graph.

D) All of the above

Answer:

A) Join operators do indeed add data to graphs and produce new graphs.
B) Structural operators operate on the graph's structure and can create new graphs.
C) Property operators modify vertex or edge properties using user-defined functions, producing new graphs.

D) Correct - All of the above statements are true.

Explanation:

Each type of operator serves to extend the capabilities of graph processing in GraphX, enabling various transformations and manipulations of graph data.

7.Which RDD operator would you use to combine two RDDs by aligning their keys and producing a new RDD with tuples of corresponding values?

A) union

B) join

C) sample

D) partitionBy

Answer:

A) Incorrect - union.
Explanation: The union operator combines two RDDs without regard to key relationships; it simply appends the elements of both RDDs.

B) Correct - join.
Explanation: The join operator is specifically designed for combining two RDDs based on their keys, resulting in an RDD of key-value pairs where the keys are aligned.

C) Incorrect - sample.
Explanation: The sample operator creates a new RDD by taking a random sample from an existing RDD, without combining two RDDs.

D) Incorrect - partitionBy.
Explanation: The partitionBy operator is used to control how data is partitioned across nodes, not for combining RDDs.

8. Which of the following is a primary benefit of using graph-based methods in data mining and machine learning?

A) Reducing the dimensionality of the data

B) Identifying influential people and information, and finding communities

C) Improving the speed of data retrieval from databases

D) Enhancing the accuracy of linear regression models

Answer:

A) Incorrect - Reducing the dimensionality of the data.
Explanation: While graph methods can assist in dimensionality reduction, techniques like PCA are more directly aimed at this task.

B) Correct - Identifying influential people and information, and finding communities.
Explanation: Graph-based methods excel in analyzing relationships and interactions within data. They help identify key players in a network and can reveal clusters or communities based on connectivity.

C) Incorrect - Improving the speed of data retrieval from databases.
Explanation: Graph methods are not primarily focused on data retrieval speed; database indexing is more relevant for that purpose.

D) Incorrect - Enhancing the accuracy of linear regression models.
Explanation: Graph-based methods are not designed to specifically improve the accuracy of linear regression; they focus on relationships in the data.

9.Which of the following accurately describes a strategy used to optimize graph computations in distributed systems?

A) Recasting graph systems optimizations as distributed join optimization and incremental materialized maintenance

B) Encoding graphs as simple arrays and using linear algebra operations

C) Expressing graph computation in sequential algorithms and optimizing with single-node processing

D) Implementing graph algorithms using recursive function calls and minimizing parallelism

Answer:
A) Correct - Recasting graph systems optimizations as distributed join optimization and incremental materialized maintenance.
Explanation: Treating graph operations similarly to joins allows for better optimization strategies, which can reduce computational overhead and improve efficiency.

B) Incorrect - Encoding graphs as simple arrays and using linear algebra operations.
Explanation: While some algorithms may leverage linear algebra, this is not a universal strategy for optimizing graph computations.

C) Incorrect - Expressing graph computation in sequential algorithms and optimizing with single-node processing.
Explanation: Distributed systems benefit from parallel processing; sequential approaches do not effectively utilize the capabilities of distributed systems.

D) Incorrect - Implementing graph algorithms using recursive function calls and minimizing parallelism.
Explanation: Recursive calls can be inefficient for large datasets; parallelism is crucial for effective distributed graph processing.

10. What are the defining traits of a Parameter Server in distributed machine learning?

S1: Distributes a model over multiple machines.

S2: It offers two operations:

(i) Pull for query parts of the model

(ii) Push for update parts of the model.

A) Only S1 is true.

B) Only S2 is true.

C) Both S1 and S2 are true.

D) NeitherS1 nor S2 are true.

The correct answer is: C) Both S1 and S2 are true.

A Parameter Server in distributed machine learning:

- **Distributes a model over multiple machines:** This allows for efficient training of large models on clusters of machines.
- **Offers two operations:**
    - **Pull:** Workers can query parts of the model from the Parameter Server.
    - **Push:** Workers can update parts of the model by pushing their computed gradients to the Parameter Server.

These two operations are essential for the coordination and synchronization of distributed machine learning algorithms.