***Hadoop Administration – PG-DHPCSA***

**Session 1**

**Introduction to Big Data**

**o What is Big Data,**

**o Big Deal about Big Data,**

**o Big Data Sources,**

**o Industries using Big Data,**

**o Big Data challenges**

**Session 2**

**Big Data Technologies and Hadoop**

**o Solution to Big Data problems,**

**o Various Big Data Technologies,**

**o Big Data/Hadoop Platforms,**

**o Hadoop Distributions and Vendors,**

**o Big Data Suites.**

**Session 3**

**Introduction to Hadoop**

**o A Brief History of Hadoop,**

**o Evolution of Hadoop,**

**o Comparison with Other Systems,**

**o Hadoop Releases**

**Session: 4**

**Hadoop Architecture**

**o Hadoop Architecture,**

**o Core components of Hadoop,**

**Session 5**

**Getting Started: Hadoop Installation**

**o Setting up a Hadoop Cluster,**

**o Logging configuration**

**o Cluster specification,**

**o Cluster Setup and Installation,**

**o Common Hadoop Shell commands**

**o Clustering Monitoring**

**o Single and Multi-Node Cluster Setup on Virtual Machine,**

**o Hadoop Configuration, Security in Hadoop, Administering Hadoop,**

**o HDFS – Monitoring & Maintenance, Hadoop benchmarks**

**o Hadoop in the cloud.**

**Session: 6**

**Hadoop Distributed File System (HDFS)**

**o Distributed File System,**

**o What is HDFS,**

**o Major goals of HDFS Design**

**o Where does HDFS fit in,**

**o Core components of HDFS,**

**o Hadoop Server Roles: Name Node, Secondary Name Node, and Data Node**

**Lab-Assignment:**

**o Run the HDFS commands, and add a one-liner understanding for each of the command.**

**o Execute the provided code using HDFS, step run and understand**

**Session: 7**

**HDFS Architecture**

**o HDFS Architecture,**

**o Scaling and Rebalancing,**

**o Big Deal about HDFS,**

**o Replication,**

**o Rack Awareness,**

**o Data Pipelining,**

**o Node Failure Management.**

**o HDFS NameNode High Availability**

**o Components and daemon of an HDFS HA-Quorum cluster**

**o HDFS Federation use case**

**o Kerberos: Role of HDFS security**

**Session 8**

**HDFS Data Storage Process**

**o HDFS Data storage process,**

**o Anatomy of writing and reading file in HDFS,**

**o HDFS user and admin commands,**

**o HDFS Web Interface.**

**Session: 9**

**Getting in touch with Map Reduce Framework**

**o Hadoop Map Reduce paradigm,**

**o Stages of MapReduce**

**o Map and Reduce tasks,**

**o Map Reduce Execution Framework,**

**o Anatomy of a Map Reduce Job run**

**Session: 10**

**YARN**

**o YARN Architecture**

**o YARN Resource Management**

**o Hadoop Schedulers**

**o Upgrading cluster from Hadoop1 to Hadoop2**

**o MapReduce job workflow on YARN**

**o Migration from MRv1 to MRv2 on YARN : Configuration changes in files**

**Session: 11**

**Security in Hadoop**

**o HDFS Security Model**

**o LDAP and Hadoop**

**o LDAP support in Hadoop**

**Session: 12,**

**Hadoop Cluster Planning**

**o Choosing hardware and operating systems,**

**o OS comparison based on features like kernel tuning, disk swapping & etc.**

**o Based on scenario and workload identify hardware, cluster size**

**o Based on scenario identify eco-system components**

**o Identify key network components, Network topology/design based on network usage in Hadoop**

**Session: 13,**

**Cluster Maintenance**

**o Managing Hadoop Process both with script and manually**

**o HDFS Maintenance tasks - Adding, decommissioning data node & etc.**

**o MapReduce Maintenance tasks - Adding, decommissioning Taskt**

**o racker, killing job/task & etc.**

**o Backup & Recovery**

**Session1**

Big Data Technologies and Hadoop

o Solution to Big Data problems,

o Various Big Data Technologies,

o Big Data/Hadoop Platforms,

o Hadoop Distributions and Vendors,

o Big Data Suites.

### **Introduction to Big Data - MCQs**

#### **Easy Questions**

1. **What is the primary characteristic of Big Data?**
   * A) Small in volume
   * B) Large in volume
   * C) Low in variety
   * D) Unstructured
   * **Answer: B) Large in volume**
2. **Which of the following is NOT typically considered a component of Big Data?**
   * A) Volume
   * B) Variety
   * C) Virtuality
   * D) Velocity
   * **Answer: C) Virtuality**
3. **Which of the following industries commonly uses Big Data?**
   * A) Retail
   * B) Healthcare
   * C) Finance
   * D) All of the above
   * **Answer: D) All of the above**
4. **Big Data refers to datasets that are too large to be processed using traditional data processing tools. What is this typically known as?**
   * A) Normal Data
   * B) Structured Data
   * C) Unstructured Data
   * D) Data Analytics
   * **Answer: C) Unstructured Data**
5. **Which of the following is a source of Big Data?**
   * A) Social media posts
   * B) Website traffic logs
   * C) Sensors and IoT devices
   * D) All of the above
   * **Answer: D) All of the above**
6. **What does the term “Big Data” commonly refer to?**
   * A) Data that is only qualitative
   * B) Data that cannot be analyzed
   * C) Extremely large datasets that require specialized tools for processing
   * D) Simple data used for decision-making
   * **Answer: C) Extremely large datasets that require specialized tools for processing**
7. **Which of the following is NOT a challenge of Big Data?**
   * A) Scalability
   * B) Data cleaning
   * C) Limited storage
   * D) Lack of variety
   * **Answer: D) Lack of variety**
8. **Which type of data is often involved in Big Data analytics?**
   * A) Structured data only
   * B) Unstructured data only
   * C) Semi-structured data only
   * D) Structured, unstructured, and semi-structured data
   * **Answer: D) Structured, unstructured, and semi-structured data**
9. **What is one of the biggest challenges organizations face when dealing with Big Data?**
   * A) Data redundancy
   * B) Data integration
   * C) Data visualization
   * D) Data categorization
   * **Answer: B) Data integration**
10. **Which term refers to the speed at which Big Data is generated and processed?**
    * A) Volume
    * B) Variety
    * C) Velocity
    * D) Value
    * **Answer: C) Velocity**

#### **Intermediate Questions**

1. **Which of the following is a key benefit of Big Data in business?**
   * A) Higher operational costs
   * B) Enhanced decision-making capabilities
   * C) Reduced amount of data storage
   * D) Decreased customer interaction
   * **Answer: B) Enhanced decision-making capabilities**
2. **What does the 'Variety' aspect of Big Data refer to?**
   * A) The volume of data
   * B) The different types and formats of data
   * C) The geographical spread of data
   * D) The speed of data generation
   * **Answer: B) The different types and formats of data**
3. **Which of the following industries has used Big Data for predictive analytics to prevent equipment failure in manufacturing?**
   * A) Agriculture
   * B) Retail
   * C) Healthcare
   * D) Manufacturing
   * **Answer: D) Manufacturing**
4. **Which of the following is an example of a source of Big Data in the healthcare industry?**
   * A) Medical records
   * B) Patient monitoring systems
   * C) Genetic data
   * D) All of the above
   * **Answer: D) All of the above**
5. **Which of the following Big Data processing frameworks is known for handling unstructured data?**
   * A) Hadoop
   * B) SQL databases
   * C) Microsoft Excel
   * D) Oracle
   * **Answer: A) Hadoop**
6. **Which of the following is an example of a Big Data challenge related to privacy?**
   * A) Data encryption
   * B) Storage capacity
   * C) Securing personal data in a decentralized system
   * D) Data variety
   * **Answer: C) Securing personal data in a decentralized system**
7. **What is the main reason social media platforms like Facebook and Twitter generate Big Data?**
   * A) User-generated content
   * B) Transaction processing
   * C) Sensor-based data
   * D) Data from external databases
   * **Answer: A) User-generated content**
8. **Which technology is commonly used for managing large datasets in Big Data environments?**
   * A) Relational Databases
   * B) NoSQL databases
   * C) Blockchain
   * D) Spreadsheets
   * **Answer: B) NoSQL databases**
9. **Which of the following is a key challenge when integrating Big Data into decision-making?**
   * A) Lack of data storage
   * B) Lack of skilled personnel
   * C) High cost of hardware
   * D) Lack of user-friendly interfaces
   * **Answer: B) Lack of skilled personnel**
10. **Which term refers to the ability to generate valuable insights from Big Data?**
    * A) Data integration
    * B) Data visualization
    * C) Data mining
    * D) Data storage
    * **Answer: C) Data mining**

#### **Difficult Questions**

1. **What is the primary role of machine learning in Big Data analytics?**
   * A) Automating data cleaning processes
   * B) Structuring unstructured data
   * C) Discovering patterns and trends in large datasets
   * D) Compressing large datasets
   * **Answer: C) Discovering patterns and trends in large datasets**
2. **Which of the following is a limitation of traditional relational databases in Big Data applications?**
   * A) High scalability
   * B) Handling unstructured data
   * C) Data integrity enforcement
   * D) Real-time processing
   * **Answer: B) Handling unstructured data**
3. **Which Big Data platform is designed for real-time data processing?**
   * A) Hadoop
   * B) Apache Spark
   * C) MongoDB
   * D) MySQL
   * **Answer: B) Apache Spark**
4. **Which of the following is a common use case for Big Data in the finance industry?**
   * A) Customer sentiment analysis
   * B) Credit scoring and fraud detection
   * C) Predictive maintenance
   * D) Supply chain optimization
   * **Answer: B) Credit scoring and fraud detection**
5. **Which of the following is a technique used for cleaning Big Data?**
   * A) Data normalization
   * B) Data anonymization
   * C) Data visualization
   * D) Data imputation
   * **Answer: D) Data imputation**
6. **Which of the following challenges is related to the 'Value' aspect of Big Data?**
   * A) Determining the relevance of data
   * B) Storage limitations
   * C) Data security
   * D) Data privacy
   * **Answer: A) Determining the relevance of data**
7. **In the context of Big Data, what does “Data Lake” refer to?**
   * A) A small-scale database
   * B) A system used to store structured data only
   * C) A storage repository that holds raw data in its native format
   * D) A method for encrypting sensitive data
   * **Answer: C) A storage repository that holds raw data in its native format**
8. **What is the biggest challenge with the integration of Big Data from different sources?**
   * A) Data inconsistency
   * B) High cost of processing
   * C) Lack of storage space
   * D) Low-quality data
   * **Answer: A) Data inconsistency**
9. **Which of the following Big Data technologies is widely used for distributed data storage?**
   * A) Hadoop Distributed File System (HDFS)
   * B) MySQL
   * C) Microsoft Excel
   * D) MongoDB
   * **Answer: A) Hadoop Distributed File System (HDFS)**
10. **Which of the following is the key challenge when using Big Data for healthcare applications?**
    * A) Lack of computing power
    * B) Privacy and security concerns
    * C) Limited data storage
    * D) Data visualization techniques
    * **Answer: B) Privacy and security concerns**

#### **Advanced Questions**

1. **Which machine learning algorithm is commonly used for clustering data in Big Data applications?**
   * A) Decision Trees
   * B) K-Means
   * C) Linear Regression
   * D) Random Forest
   * **Answer: B) K-Means**
2. **Which of the following is a disadvantage of using cloud platforms for storing Big Data?**
   * A) High data transfer rates
   * B) Security and privacy concerns
   * C) Limited computational power
   * D) Easy integration with other platforms
   * **Answer: B) Security and privacy concerns**
3. **Which of the following is true about NoSQL databases in Big Data?**
   * A) They are designed for handling structured data only
   * B) They are more flexible and scalable than traditional relational databases
   * C) They require SQL for querying
   * D) They do not support distributed computing
   * **Answer: B) They are more flexible and scalable than traditional relational databases**
4. **In Big Data analytics, what is a 'data silo'?**
   * A) A central repository for all data
   * B) Isolated data that cannot be accessed across departments
   * C) A data format for machine learning
   * D) A tool for data cleaning
   * **Answer: B) Isolated data that cannot be accessed across departments**
5. **Which of the following techniques is used to deal with the ‘velocity’ challenge of Big Data?**
   * A) Data warehousing
   * B) Stream processing
   * C) Data normalization
   * D) Data imputation
   * **Answer: B) Stream processing**
6. **Which of the following is an advantage of using Big Data in the retail industry?**
   * A) Higher product prices
   * B) Improved customer segmentation and targeting
   * C) Reduced need for data security
   * D) Fewer customer complaints
   * **Answer: B) Improved customer segmentation and targeting**
7. **Which of the following is a key advantage of cloud computing in handling Big Data?**
   * A) Better data integration
   * B) Flexible and scalable storage and processing power
   * C) Data compression
   * D) Increased data privacy
   * **Answer: B) Flexible and scalable storage and processing power**
8. **Which Big Data technology is known for processing large datasets in parallel across multiple machines?**
   * A) Hadoop
   * B) Spark
   * C) TensorFlow
   * D) SQL Server
   * **Answer: A) Hadoop**
9. **Which of the following best describes the 'veracity' aspect of Big Data?**
   * A) The volume of data
   * B) The speed of data
   * C) The accuracy and reliability of data
   * D) The variety of data
   * **Answer: C) The accuracy and reliability of data**
10. **What is one of the core advantages of using Big Data in predictive analytics?**
    * A) Improved data compression
    * B) More accurate forecasting
    * C) Reduced need for data security
    * D) Increased data redundancy
    * **Answer: B) More accurate forecasting**

#### **Intermediate to Advanced Questions**

1. **Which of the following is a key benefit of using Hadoop for Big Data storage and processing?**

* A) Supports only structured data
* B) Can handle massive amounts of unstructured data across distributed systems
* C) Requires centralized processing
* D) Relies on SQL for querying
* **Answer: B) Can handle massive amounts of unstructured data across distributed systems**

1. **What is the primary purpose of data governance in the context of Big Data?**

* A) Ensuring data is stored securely in the cloud
* B) Improving the speed of data processing
* C) Managing data quality, security, and compliance across the organization
* D) Automating the collection of data from external sources
* **Answer: C) Managing data quality, security, and compliance across the organization**

1. **Which of the following best describes 'Real-time Big Data Analytics'?**

* A) Analyzing data after it has been stored for a long period
* B) Analyzing data as it is generated, with minimal delay
* C) Analyzing data in batch processes
* D) Storing data without any analysis
* **Answer: B) Analyzing data as it is generated, with minimal delay**

1. **Which of the following is an example of Big Data used in the transportation industry?**

* A) Real-time traffic monitoring systems
* B) Financial fraud detection
* C) E-commerce customer reviews
* D) Agricultural yield prediction
* **Answer: A) Real-time traffic monitoring systems**

1. **What is the main purpose of a 'Data Warehouse' in Big Data architecture?**

* A) To store raw, unprocessed data in its native format
* B) To store structured data from different sources for analysis and reporting
* C) To provide real-time processing of transactional data
* D) To visualize data trends
* **Answer: B) To store structured data from different sources for analysis and reporting**

1. **Which of the following describes 'MapReduce' in Big Data processing?**

* A) A data visualization tool used for Big Data
* B) A machine learning algorithm for Big Data
* C) A distributed processing technique that divides tasks into smaller sub-tasks
* D) A NoSQL database used to store Big Data
* **Answer: C) A distributed processing technique that divides tasks into smaller sub-tasks**

1. **Which of the following is an example of 'Semi-Structured Data' in Big Data?**

* A) Relational databases
* B) Sensor data in CSV files
* C) Data in XML or JSON format
* D) Traditional text documents
* **Answer: C) Data in XML or JSON format**

1. **Which of the following tools is commonly used for real-time Big Data stream processing?**

* A) Hadoop
* B) Apache Kafka
* C) MySQL
* D) Oracle DB
* **Answer: B) Apache Kafka**

1. **What is a key challenge when using Big Data for predictive analytics?**

* A) Lack of data storage space
* B) Ensuring data privacy and compliance with regulations
* C) The inability to store unstructured data
* D) Limited computational power for analysis
* **Answer: B) Ensuring data privacy and compliance with regulations**

1. **Which of the following best describes the role of 'Data Scientists' in Big Data analytics?**

* A) Collecting data from various sources
* B) Storing data securely in cloud platforms
* C) Analyzing large datasets using algorithms and statistical methods to derive actionable insights
* D) Managing the physical infrastructure for data storage
* **Answer: C) Analyzing large datasets using algorithms and statistical methods to derive actionable insights**

###########################################################################

**Session 2**

**Big Data Technologies and Hadoop**

o Solution to Big Data problems,

o Various Big Data Technologies,

o Big Data/Hadoop Platforms,

o Hadoop Distributions and Vendors,

o Big Data Suites.

### **Easy-Level Questions (1-15)**

1. **What is Big Data?**
   * A) Data that is large in size
   * B) Data that is difficult to manage
   * C) Data that is high in volume, velocity, and variety
   * D) All of the above
   * **Answer:** D) All of the above
2. **Which of the following tools is specifically used for storing Big Data?**
   * A) MySQL
   * B) MongoDB
   * C) Hadoop Distributed File System (HDFS)
   * D) Oracle
   * **Answer:** C) Hadoop Distributed File System (HDFS)
3. **Which company originally developed Hadoop?**
   * A) Google
   * B) Yahoo!
   * C) Facebook
   * D) Amazon
   * **Answer:** B) Yahoo!
4. **Hadoop is primarily written in which programming language?**
   * A) Java
   * B) Python
   * C) Scala
   * D) C++
   * **Answer:** A) Java
5. **Which of these is NOT part of the Hadoop ecosystem?**
   * A) HDFS
   * B) MapReduce
   * C) Hive
   * D) Spark
   * **Answer:** D) Spark
6. **Which of the following is a Hadoop distribution vendor?**
   * A) MongoDB
   * B) Cloudera
   * C) Microsoft
   * D) Oracle
   * **Answer:** B) Cloudera
7. **In Hadoop, what is the default file system?**
   * A) HDFS
   * B) NTFS
   * C) ext4
   * D) FAT32
   * **Answer:** A) HDFS
8. **What is the purpose of the NameNode in Hadoop?**
   * A) Managing task execution
   * B) Storing the actual data
   * C) Managing file metadata
   * D) Processing data
   * **Answer:** C) Managing file metadata
9. **What does the "YARN" component in Hadoop do?**
   * A) Stores data
   * B) Provides a resource management layer
   * C) Handles the data processing logic
   * D) Manages the file system
   * **Answer:** B) Provides a resource management layer
10. **Which of the following is a tool used for data querying in the Hadoop ecosystem?**
    * A) Pig
    * B) Hadoop Streaming
    * C) Hive
    * D) HBase
    * **Answer:** C) Hive
11. **What does "MapReduce" refer to in Hadoop?**
    * A) Data storage format
    * B) A method of distributed data processing
    * C) A file format
    * D) A data compression technique
    * **Answer:** B) A method of distributed data processing
12. **Which of the following is NOT a key component of the Hadoop ecosystem?**
    * A) HBase
    * B) Spark
    * C) SQL Server
    * D) Pig
    * **Answer:** C) SQL Server
13. **Which of these is a NoSQL database used with Big Data?**
    * A) SQL Server
    * B) HBase
    * C) Oracle
    * D) MySQL
    * **Answer:** B) HBase
14. **Which of the following is the primary benefit of Big Data technologies?**
    * A) Faster data processing
    * B) Handling data of varying sizes and types
    * C) Real-time analytics
    * D) All of the above
    * **Answer:** D) All of the above

### **Intermediate-Level Questions (16-35)**

1. **What is the role of the DataNode in Hadoop?**
   * A) Stores the metadata
   * B) Manages the file system
   * C) Stores the actual data
   * D) Runs the MapReduce jobs
   * **Answer:** C) Stores the actual data
2. **Which Hadoop distribution is known for providing a comprehensive suite of services and tools for Big Data?**
   * A) Cloudera
   * B) Hortonworks
   * C) MapR
   * D) All of the above
   * **Answer:** D) All of the above
3. **Which Big Data technology is primarily used for real-time data processing?**
   * A) Hadoop
   * B) Apache Kafka
   * C) Apache Spark
   * D) Pig
   * **Answer:** C) Apache Spark
4. **In Hadoop, what is the purpose of the Secondary NameNode?**
   * A) Stores the data
   * B) Provides backups of the NameNode
   * C) Handles job scheduling
   * D) Processes data in parallel
   * **Answer:** B) Provides backups of the NameNode
5. **What is the primary function of Apache HBase?**
   * A) Relational database management
   * B) Distributed file system
   * C) Column-family NoSQL database
   * D) Data analysis tool
   * **Answer:** C) Column-family NoSQL database
6. **Which Hadoop component allows users to write data in a more user-friendly way using SQL-like queries?**
   * A) MapReduce
   * B) Hive
   * C) Pig
   * D) YARN
   * **Answer:** B) Hive
7. **Which of these is an open-source alternative to Hadoop for distributed data processing?**
   * A) Apache Kafka
   * B) Apache Storm
   * C) Apache Spark
   * D) SQL Server
   * **Answer:** C) Apache Spark
8. **Which of the following is true about Apache Flume?**
   * A) It is used for data processing
   * B) It is a data collection tool for streaming data
   * C) It is a storage layer for Big Data
   * D) It is used for data visualization
   * **Answer:** B) It is a data collection tool for streaming data
9. **Which component of the Hadoop ecosystem helps to schedule and manage jobs across the cluster?**
   * A) MapReduce
   * B) YARN
   * C) Pig
   * D) Hive
   * **Answer:** B) YARN
10. **Which is the major difference between Hadoop MapReduce and Apache Spark?**
    * A) Spark is faster because it stores data in memory
    * B) MapReduce is faster than Spark
    * C) Spark is not open-source
    * D) Spark does not support batch processing
    * **Answer:** A) Spark is faster because it stores data in memory
11. **Which of these is a part of the Big Data Suite used for distributed storage and processing?**
    * A) Apache Kafka
    * B) Hadoop
    * C) MongoDB
    * D) All of the above
    * **Answer:** D) All of the above
12. **Which Hadoop distribution is known for its focus on the open-source Apache projects and compatibility?**
    * A) Cloudera
    * B) Hortonworks
    * C) MapR
    * D) Microsoft Azure
    * **Answer:** B) Hortonworks
13. **What is the key difference between Hadoop 1.x and Hadoop 2.x?**
    * A) Hadoop 2.x uses YARN for resource management
    * B) Hadoop 1.x has better scalability
    * C) Hadoop 2.x is slower
    * D) Hadoop 1.x supports real-time processing
    * **Answer:** A) Hadoop 2.x uses YARN for resource management
14. **Which of the following Big Data technologies is specifically designed for in-memory processing?**
    * A) Hadoop
    * B) Apache Flume
    * C) Apache Spark
    * D) Apache Kafka
    * **Answer:** C) Apache Spark
15. **Which of the following is a key feature of Apache Kafka?**
    * A) Distributed stream processing
    * B) Data storage system
    * C) Data querying engine
    * D) Batch processing engine
    * **Answer:** A) Distributed stream processing
16. **What is the purpose of Apache Zookeeper in the Hadoop ecosystem?**
    * A) Data querying tool
    * B) Managing distributed coordination and synchronization
    * C) Data storage system
    * D) In-memory computation
    * **Answer:** B) Managing distributed coordination and synchronization
17. **Which of the following tools is used for data processing in a more flexible, scripting-style manner in the Hadoop ecosystem?**
    * A) Pig
    * B) Hive
    * C) MapReduce
    * D) HBase
    * **Answer:** A) Pig
18. **Which of the following Hadoop components helps in fault tolerance by duplicating data blocks?**
    * A) HDFS
    * B) Pig
    * C) YARN
    * D) Hive
    * **Answer:** A) HDFS
19. **Which of the following is a popular Hadoop distribution provided by Amazon?**
    * A) Amazon EMR
    * B) Cloudera
    * C) MapR
    * D) Hortonworks
    * **Answer:** A) Amazon EMR
20. **Which is the storage format that Hadoop uses to store large datasets in a compressed and optimized format?**
    * A) Parquet
    * B) CSV
    * C) XML
    * D) JSON
    * **Answer:** A) Parquet

### **Hard-Level Questions (36-50)**

1. **What is the main advantage of using Hadoop Distributed File System (HDFS)?**
   * A) High availability
   * B) Fault tolerance
   * C) High throughput
   * D) All of the above
   * **Answer:** D) All of the above
2. **Which Hadoop distribution provides a unified platform for managing Big Data workloads and is used by many large enterprises?**
   * A) MapR
   * B) Cloudera
   * C) Amazon EMR
   * D) Microsoft HDInsight
   * **Answer:** B) Cloudera
3. **What is the role of the ResourceManager in Hadoop YARN?**
   * A) Allocates resources for jobs
   * B) Stores the job metadata
   * C) Processes the data
   * D) Manages the file system
   * **Answer:** A) Allocates resources for jobs
4. **Which of the following Hadoop components uses the concept of "Data Locality" to optimize performance?**
   * A) HDFS
   * B) YARN
   * C) MapReduce
   * D) Hive
   * **Answer:** C) MapReduce
5. **What is the typical replication factor for HDFS blocks in Hadoop?**
   * A) 1
   * B) 2
   * C) 3
   * D) 5
   * **Answer:** C) 3
6. **What is the function of the "Client" in a Hadoop ecosystem?**
   * A) Initiates the job execution
   * B) Stores the actual data
   * C) Manages cluster resources
   * D) Coordinates the replication of data blocks
   * **Answer:** A) Initiates the job execution
7. **In Hadoop, how is data partitioned across the cluster?**
   * A) By a primary key
   * B) By using sharding techniques
   * C) By dividing into blocks
   * D) By columns
   * **Answer:** C) By dividing into blocks
8. **What is the key difference between MapReduce and Spark in terms of fault tolerance?**
   * A) Spark supports real-time processing, MapReduce does not
   * B) MapReduce stores intermediate data on disk, while Spark stores it in memory
   * C) Spark does not offer fault tolerance
   * D) MapReduce supports in-memory computations, Spark does not
   * **Answer:** B) MapReduce stores intermediate data on disk, while Spark stores it in memory
9. **What does the acronym "YARN" stand for in Hadoop?**
   * A) Yet Another Resource Negotiator
   * B) Yet Another Resource Node
   * C) Yarn Allocation and Resource Node
   * D) YARN Application Resource Node
   * **Answer:** A) Yet Another Resource Negotiator
10. **Which of the following Big Data technologies can be used for batch processing?**
    * A) Apache Flume
    * B) Apache Spark
    * C) Apache Storm
    * D) Apache Kafka
    * **Answer:** B) Apache Spark
11. **Which of these tools is commonly used for real-time data streaming in Big Data environments?**
    * A) Apache Kafka
    * B) Apache Spark
    * C) Apache Hive
    * D) Apache HBase
    * **Answer:** A) Apache Kafka
12. **In which scenario is HBase most commonly used in the Hadoop ecosystem?**
    * A) Storing data in a relational database
    * B) Storing data in a column-family NoSQL database
    * C) Performing in-memory data processing
    * D) Streaming data to external sources
    * **Answer:** B) Storing data in a column-family NoSQL database
13. **Which of these is true about the MapReduce framework?**
    * A) It allows for both batch and real-time processing
    * B) It is not scalable
    * C) It processes data in parallel across a distributed system
    * D) It uses a centralized server to process data
    * **Answer:** C) It processes data in parallel across a distributed system
14. **Which of the following Big Data tools provides a way to convert structured data into an unstructured format for processing?**
    * A) Apache Hive
    * B) Apache Pig
    * C) Apache Spark
    * D) Apache Kafka
    * **Answer:** B) Apache Pig
15. **What is the main advantage of Hadoop over traditional data processing frameworks?**
    * A) Scalability to process large datasets
    * B) Real-time processing capabilities
    * C) High-cost efficiency
    * D) Centralized data storage system
    * **Answer:** A) Scalability to process large datasets

###########################################################################

**Session3:**

**Introduction to Hadoop**

o A Brief History of Hadoop,

o Evolution of Hadoop,

o Comparison with Other Systems,

o Hadoop Releases

### **Easy-Level Questions (1-15)**

1. **Who is credited with the creation of Hadoop?**
   * A) Doug Cutting
   * B) Mark Zuckerberg
   * C) Larry Page
   * D) Jeff Bezos
   * **Answer:** A) Doug Cutting
2. **In which year was Hadoop created?**
   * A) 2000
   * B) 2005
   * C) 2007
   * D) 2010
   * **Answer:** B) 2005
3. **What is the original purpose of Hadoop?**
   * A) Web search
   * B) Distributed storage and processing of Big Data
   * C) Social media platform
   * D) Cloud computing
   * **Answer:** B) Distributed storage and processing of Big Data
4. **Hadoop is primarily designed to work with what type of data?**
   * A) Relational data
   * B) Semi-structured data
   * C) Unstructured data
   * D) All types of data
   * **Answer:** D) All types of data
5. **Which of these was the main inspiration for Hadoop?**
   * A) Amazon's Dynamo
   * B) Google File System (GFS)
   * C) Facebook's graph database
   * D) Oracle databases
   * **Answer:** B) Google File System (GFS)
6. **What was Hadoop originally called?**
   * A) Nutch
   * B) HDFS
   * C) Hive
   * D) Spark
   * **Answer:** A) Nutch
7. **Which of the following is a key feature of Hadoop?**
   * A) Centralized architecture
   * B) Scalability
   * C) Limited data processing
   * D) High latency
   * **Answer:** B) Scalability
8. **What is the Hadoop Distributed File System (HDFS) used for?**
   * A) Storing metadata
   * B) Storing data across multiple machines
   * C) Processing large data sets
   * D) Data visualization
   * **Answer:** B) Storing data across multiple machines
9. **Which company first contributed significantly to the Hadoop project?**
   * A) Microsoft
   * B) Yahoo!
   * C) IBM
   * D) Amazon
   * **Answer:** B) Yahoo!
10. **Which of the following is a major component of Hadoop?**
    * A) MapReduce
    * B) Hive
    * C) Pig
    * D) All of the above
    * **Answer:** D) All of the above
11. **Which of these is NOT a core component of Hadoop?**
    * A) HDFS
    * B) YARN
    * C) MongoDB
    * D) MapReduce
    * **Answer:** C) MongoDB
12. **Which Hadoop component is responsible for job scheduling and resource management?**
    * A) YARN
    * B) MapReduce
    * C) HDFS
    * D) Hive
    * **Answer:** A) YARN
13. **What is the main purpose of the MapReduce component in Hadoop?**
    * A) Storing data
    * B) Resource management
    * C) Processing data in parallel
    * D) Querying data
    * **Answer:** C) Processing data in parallel
14. **Which of the following systems is Hadoop compared with when it comes to Big Data processing?**
    * A) Oracle Database
    * B) SQL Server
    * C) Spark
    * D) MySQL
    * **Answer:** C) Spark

### **Intermediate-Level Questions (16-35)**

1. **Which system did Hadoop initially aim to mimic or improve upon?**
   * A) Amazon's EC2
   * B) Google's MapReduce
   * C) IBM's DB2
   * D) Microsoft's SQL Server
   * **Answer:** B) Google's MapReduce
2. **What was the first major release of Hadoop?**
   * A) Hadoop 0.1.0
   * B) Hadoop 1.0
   * C) Hadoop 0.20.0
   * D) Hadoop 2.0
   * **Answer:** A) Hadoop 0.1.0
3. **What major change was introduced in Hadoop 2.x compared to Hadoop 1.x?**
   * A) Support for larger datasets
   * B) Introduction of YARN for resource management
   * C) A new file format
   * D) Support for real-time processing
   * **Answer:** B) Introduction of YARN for resource management
4. **Which Hadoop version introduced HDFS Federation to allow multiple NameNodes?**
   * A) Hadoop 1.0
   * B) Hadoop 2.0
   * C) Hadoop 3.0
   * D) Hadoop 4.0
   * **Answer:** B) Hadoop 2.0
5. **Which of the following is a key differentiator between Hadoop and traditional relational databases?**
   * A) Hadoop uses a single server for processing
   * B) Hadoop is designed for distributed, parallel processing
   * C) Hadoop only supports structured data
   * D) Hadoop has high query optimization capabilities
   * **Answer:** B) Hadoop is designed for distributed, parallel processing
6. **Which of the following is the major benefit of Hadoop's distributed architecture?**
   * A) Low cost for large data sets
   * B) Real-time data analysis
   * C) High security for data
   * D) No need for parallel computation
   * **Answer:** A) Low cost for large data sets
7. **Which of the following systems is Hadoop most commonly compared with for batch processing?**
   * A) Apache Spark
   * B) Oracle Exadata
   * C) SQL Server
   * D) PostgreSQL
   * **Answer:** A) Apache Spark
8. **What does Hadoop’s HDFS stand for?**
   * A) Hadoop Distributed File System
   * B) High Density File System
   * C) Hadoop Data File System
   * D) High Definition File System
   * **Answer:** A) Hadoop Distributed File System
9. **What is the primary reason for the popularity of Hadoop?**
   * A) It is suitable for both small and large data processing
   * B) It can run on a single server
   * C) It is a relational database management system
   * D) It does not require specialized hardware
   * **Answer:** D) It does not require specialized hardware
10. **Which Hadoop release introduced the concept of the "ResourceManager" in YARN?**
    * A) Hadoop 1.0
    * B) Hadoop 2.0
    * C) Hadoop 3.0
    * D) Hadoop 4.0
    * **Answer:** B) Hadoop 2.0
11. **Which version of Hadoop focused on improving the scalability and performance of the system?**
    * A) Hadoop 0.x
    * B) Hadoop 1.x
    * C) Hadoop 2.x
    * D) Hadoop 3.x
    * **Answer:** C) Hadoop 2.x
12. **Which of these Hadoop components enables data querying through SQL-like syntax?**
    * A) HBase
    * B) Hive
    * C) Pig
    * D) YARN
    * **Answer:** B) Hive
13. **Which of the following is NOT a version of Hadoop?**
    * A) Hadoop 1.x
    * B) Hadoop 2.x
    * C) Hadoop 3.x
    * D) Hadoop 5.x
    * **Answer:** D) Hadoop 5.x
14. **In which Hadoop release was the ability to store data in multiple clusters added?**
    * A) Hadoop 0.20.0
    * B) Hadoop 1.0
    * C) Hadoop 2.x
    * D) Hadoop 3.x
    * **Answer:** C) Hadoop 2.x
15. **Which of the following Hadoop components provides a SQL-like interface to interact with data?**
    * A) MapReduce
    * B) Hive
    * C) Pig
    * D) YARN
    * **Answer:** B) Hive
16. **Which of the following Hadoop versions provided improvements in workload management and fault tolerance?**
    * A) Hadoop 1.0
    * B) Hadoop 2.0
    * C) Hadoop 3.0
    * D) Hadoop 4.0
    * **Answer:** B) Hadoop 2.0

### **Hard-Level Questions (36-50)**

1. **Which of the following is the primary reason Hadoop outperforms traditional relational databases for Big Data processing?**
   * A) Hadoop’s parallel processing capability
   * B) Hadoop uses a centralized architecture
   * C) Hadoop supports ACID transactions
   * D) Hadoop is a fully managed service
   * **Answer:** A) Hadoop’s parallel processing capability
2. **Which release of Hadoop introduced improvements related to HDFS storage efficiency and reduced block replication?**
   * A) Hadoop 1.0
   * B) Hadoop 2.0
   * C) Hadoop 3.0
   * D) Hadoop 4.0
   * **Answer:** C) Hadoop 3.0
3. **Hadoop's evolution has led to several new systems and tools. Which of these is a newer alternative to MapReduce for in-memory processing?**
   * A) Apache Hive
   * B) Apache Pig
   * C) Apache Spark
   * D) Apache HBase
   * **Answer:** C) Apache Spark
4. **Which of the following best describes the improvements introduced in Hadoop 3.x over earlier versions?**
   * A) Introduction of real-time processing capabilities
   * B) Enhanced resource management and fault tolerance
   * C) Focus on centralized data processing
   * D) Introduction of the ResourceManager architecture
   * **Answer:** B) Enhanced resource management and fault tolerance
5. **Which release of Hadoop introduced support for erasure coding, a more efficient method of data redundancy?**
   * A) Hadoop 2.x
   * B) Hadoop 3.x
   * C) Hadoop 1.x
   * D) Hadoop 4.x
   * **Answer:** B) Hadoop 3.x
6. **Which of these Hadoop versions is considered stable for large-scale enterprise use?**
   * A) Hadoop 0.x
   * B) Hadoop 1.x
   * C) Hadoop 2.x
   * D) Hadoop 3.x
   * **Answer:** D) Hadoop 3.x
7. **Which Hadoop version supports running on Docker containers?**
   * A) Hadoop 1.x
   * B) Hadoop 2.x
   * C) Hadoop 3.x
   * D) Hadoop 4.x
   * **Answer:** C) Hadoop 3.x
8. **In Hadoop, what is the default block size for HDFS in version 3.x?**
   * A) 64MB
   * B) 128MB
   * C) 256MB
   * D) 512MB
   * **Answer:** B) 128MB
9. **Which of the following Hadoop versions introduced support for GPUs for faster processing?**
   * A) Hadoop 2.x
   * B) Hadoop 3.x
   * C) Hadoop 1.x
   * D) Hadoop 4.x
   * **Answer:** B) Hadoop 3.x
10. **What major feature was introduced in Hadoop 2.0 to improve scalability and efficiency?**
    * A) YARN (Yet Another Resource Negotiator)
    * B) HBase
    * C) HDFS Federation
    * D) Hadoop Distributed Scheduling
    * **Answer:** A) YARN (Yet Another Resource Negotiator)
11. **Which of the following is the first step in Hadoop’s data processing pipeline?**

* A) Data is processed by the Map function
* B) Data is stored in HDFS
* C) Data is split into blocks
* D) Data is passed through a reducer
* **Answer:** B) Data is stored in HDFS

1. **Which of the following is used to run jobs on Hadoop clusters?**

* A) HDFS
* B) YARN
* C) MapReduce
* D) Hive
* **Answer:** C) MapReduce

1. **What feature of Hadoop allows data to be replicated across multiple nodes?**

* A) Data Compression
* B) Data Locality
* C) Data Replication
* D) Data Sharding
* **Answer:** C) Data Replication

1. **Which of the following is a major disadvantage of traditional RDBMS when compared to Hadoop?**

* A) Limited scalability
* B) Lack of fault tolerance
* C) Complex query language
* D) Support for structured data only
* **Answer:** A) Limited scalability

1. **Which component of Hadoop is responsible for job execution?**

* A) ResourceManager
* B) HDFS
* C) DataNode
* D) NodeManager
* **Answer:** D) NodeManager

1. **Which of the following Hadoop tools allows users to interact with large datasets using SQL-like queries?**

* A) Pig
* B) HBase
* C) Hive
* D) YARN
* **Answer:** C) Hive

1. **Hadoop's MapReduce framework is based on which computational model?**

* A) Master-slave model
* B) Shared-memory model
* C) Map and Reduce model
* D) Client-server model
* **Answer:** C) Map and Reduce model

1. **Which of the following is NOT a Hadoop ecosystem component?**

* A) HBase
* B) Pig
* C) Cassandra
* D) HDFS
* **Answer:** C) Cassandra

1. **Which of the following Hadoop tools is best used for real-time streaming data processing?**

* A) Apache Kafka
* B) Apache Spark
* C) Apache Flume
* D) Apache Storm
* **Answer:** D) Apache Storm

1. **What is the default block size in HDFS for Hadoop 2.x?**

* A) 32MB
* B) 64MB
* C) 128MB
* D) 256MB
* **Answer:** C) 128MB

1. **Which of these file formats is optimized for large-scale data storage in Hadoop?**

* A) JSON
* B) CSV
* C) Parquet
* D) XML
* **Answer:** C) Parquet

1. **What is the primary use case for Apache Pig in the Hadoop ecosystem?**

* A) Data storage
* B) Data querying
* C) Data processing with a high-level scripting language
* D) Data replication
* **Answer:** C) Data processing with a high-level scripting language

1. **Which of the following describes the function of the “NameNode” in HDFS?**

* A) Stores data blocks
* B) Manages metadata and file system namespace
* C) Processes data
* D) Provides resources for job execution
* **Answer:** B) Manages metadata and file system namespace

1. **Which Hadoop component manages the lifecycle of containers that run applications?**

* A) ResourceManager
* B) NameNode
* C) NodeManager
* D) JobTracker
* **Answer:** C) NodeManager

### **Intermediate-Level Questions (31-45)**

1. **Which of these Hadoop versions added the feature of HDFS Federation, allowing multiple NameNodes?**

* A) Hadoop 1.x
* B) Hadoop 2.x
* C) Hadoop 3.x
* D) Hadoop 4.x
* **Answer:** B) Hadoop 2.x

1. **What is the major advantage of using YARN in Hadoop?**

* A) It provides better fault tolerance
* B) It simplifies data storage
* C) It enables multi-tenancy and better resource management
* D) It increases storage capacity
* **Answer:** C) It enables multi-tenancy and better resource management

1. **Which Hadoop component allows you to process unstructured data in a parallel way using MapReduce?**

* A) Pig
* B) Hive
* C) HBase
* D) Spark
* **Answer:** A) Pig

1. **Which Hadoop release introduced the ability to store large amounts of metadata to make HDFS more scalable?**

* A) Hadoop 1.x
* B) Hadoop 2.x
* C) Hadoop 3.x
* D) Hadoop 4.x
* **Answer:** C) Hadoop 3.x

1. **Which of these is a critical feature of Hadoop 3.x compared to previous versions?**

* A) Enhanced security features
* B) Support for Docker containers
* C) Introduction of erasure coding
* D) NoSQL support
* **Answer:** C) Introduction of erasure coding

1. **Which of the following tools is used in Hadoop for data transformation and analysis using an SQL-like language?**

* A) Apache Pig
* B) Apache Hive
* C) Apache Kafka
* D) Apache HBase
* **Answer:** B) Apache Hive

1. **Which of the following is NOT a part of the Hadoop ecosystem?**

* A) HBase
* B) Pig
* C) MongoDB
* D) Hive
* **Answer:** C) MongoDB

1. **Which component of Hadoop handles the actual data storage on physical servers?**

* A) NameNode
* B) DataNode
* C) ResourceManager
* D) JobTracker
* **Answer:** B) DataNode

1. **What is a major difference between Apache Spark and Hadoop MapReduce?**

* A) Spark processes data in real-time, while MapReduce is batch-oriented
* B) Spark is slower than MapReduce
* C) Spark requires more memory than MapReduce
* D) Spark uses a central server for processing
* **Answer:** A) Spark processes data in real-time, while MapReduce is batch-oriented

1. **Which of the following versions of Hadoop introduced the YARN framework for resource management?**

* A) Hadoop 1.x
* B) Hadoop 2.x
* C) Hadoop 3.x
* D) Hadoop 4.x
* **Answer:** B) Hadoop 2.x

1. **What is the main goal of HDFS's block replication feature?**

* A) To increase the speed of data processing
* B) To ensure high availability and fault tolerance
* C) To compress data for storage
* D) To split large data sets into smaller files
* **Answer:** B) To ensure high availability and fault tolerance

1. **Which Hadoop component helps in storing structured, semi-structured, and unstructured data?**

* A) MapReduce
* B) HDFS
* C) Hive
* D) HBase
* **Answer:** B) HDFS

1. **Which of these Hadoop tools is primarily used for real-time data streaming?**

* A) Apache Kafka
* B) Apache Pig
* C) Apache HBase
* D) Apache Flume
* **Answer:** A) Apache Kafka

1. **What is a key advantage of Hadoop 3.x over Hadoop 2.x in terms of fault tolerance?**

* A) More efficient replication
* B) Support for stronger encryption
* C) Use of erasure coding for data redundancy
* D) Enhanced resource allocation
* **Answer:** C) Use of erasure coding for data redundancy

1. **Which of the following is true about Hadoop’s scalability?**

* A) Hadoop requires a very large number of nodes to function effectively
* B) Hadoop’s scalability depends on the data processing tool used (e.g., MapReduce)
* C) Hadoop can easily scale both horizontally and vertically
* D) Hadoop can only scale vertically
* **Answer:** C) Hadoop can easily scale both horizontally and vertically

### **Hard-Level Questions (46-60)**

1. **Which feature of Hadoop allows it to handle a variety of data types and sources efficiently?**

* A) Data Sharding
* B) In-memory computing
* C) Schema-on-read
* D) Centralized data processing
* **Answer:** C) Schema-on-read

1. **What is the key advantage of HDFS over traditional storage systems?**

* A) Faster data retrieval
* B) High availability and fault tolerance
* C) Simplified data management
* D) Real-time analytics
* **Answer:** B) High availability and fault tolerance

1. **Which of the following Hadoop components is most associated with providing real-time, low-latency access to large datasets?**

* A) Apache Spark
* B) Apache HBase
* C) Hadoop MapReduce
* D) Apache Kafka
* **Answer:** B) Apache HBase

1. **What is the default replication factor of HDFS in Hadoop?**

* A) 2
* B) 3
* C) 4
* D) 5
* **Answer:** B) 3

1. **What is a limitation of Hadoop’s MapReduce framework?**

* A) Lack of support for parallel processing
* B) Poor support for iterative algorithms
* C) Inability to process unstructured data
* D) Poor integration with relational databases
* **Answer:** B) Poor support for iterative algorithms

############################################################################

**Session 4** :

**Hadoop Architecture**

o Hadoop Architecture,

o Core components of Hadoop

### **Easy-Level Questions (1-10)**

1. **Which of the following is a core component of Hadoop?**
   * A) HDFS
   * B) YARN
   * C) MapReduce
   * D) All of the above
   * **Answer:** D) All of the above
2. **In Hadoop, what does HDFS stand for?**
   * A) Hadoop Distributed File System
   * B) High Definition File System
   * C) Hadoop Distributed Fault System
   * D) Hadoop Data Federation System
   * **Answer:** A) Hadoop Distributed File System
3. **Which Hadoop component is responsible for job scheduling and resource management?**
   * A) HDFS
   * B) ResourceManager
   * C) NameNode
   * D) JobTracker
   * **Answer:** B) ResourceManager
4. **What does the NameNode in HDFS do?**
   * A) Stores data blocks
   * B) Manages metadata and the namespace of the file system
   * C) Schedules jobs on the cluster
   * D) Executes MapReduce tasks
   * **Answer:** B) Manages metadata and the namespace of the file system
5. **Which of the following is the primary responsibility of a DataNode in HDFS?**
   * A) Storing data blocks
   * B) Managing job execution
   * C) Coordinating resource allocation
   * D) Providing security for data
   * **Answer:** A) Storing data blocks
6. **Which component of Hadoop is responsible for executing MapReduce tasks?**
   * A) ResourceManager
   * B) NodeManager
   * C) JobTracker
   * D) DataNode
   * **Answer:** C) JobTracker
7. **Which of the following components is responsible for resource management in YARN?**
   * A) ResourceManager
   * B) DataNode
   * C) NameNode
   * D) JobTracker
   * **Answer:** A) ResourceManager
8. **Which component stores the actual data in Hadoop’s distributed file system?**
   * A) ResourceManager
   * B) NameNode
   * C) DataNode
   * D) HBase
   * **Answer:** C) DataNode
9. **Which of the following is a job scheduling component in Hadoop 2.x?**
   * A) ResourceManager
   * B) JobTracker
   * C) TaskTracker
   * D) NodeManager
   * **Answer:** A) ResourceManager
10. **Which of these describes the function of the HDFS "Secondary NameNode"?**

* A) It stores the actual data blocks
* B) It backs up the NameNode’s metadata and file system information
* C) It manages the job execution process
* D) It coordinates MapReduce job execution
* **Answer:** B) It backs up the NameNode’s metadata and file system information

### **Intermediate-Level Questions (11-25)**

1. **In Hadoop, which of the following components helps in job execution across the cluster?**

* A) NameNode
* B) JobTracker
* C) ResourceManager
* D) DataNode
* **Answer:** B) JobTracker

1. **What is the role of YARN’s NodeManager in Hadoop?**

* A) It manages the data blocks in the HDFS
* B) It manages job execution and resource monitoring for containers
* C) It performs job scheduling across nodes
* D) It stores metadata for the HDFS
* **Answer:** B) It manages job execution and resource monitoring for containers

1. **Which of the following best describes Hadoop’s architecture?**

* A) Centralized
* B) Peer-to-peer
* C) Client-server
* D) Master-slave
* **Answer:** D) Master-slave

1. **Which Hadoop component is responsible for coordinating all the tasks in a job?**

* A) JobTracker
* B) ResourceManager
* C) NameNode
* D) DataNode
* **Answer:** A) JobTracker

1. **Which of the following can be considered a "slave" node in Hadoop?**

* A) NameNode
* B) JobTracker
* C) DataNode
* D) ResourceManager
* **Answer:** C) DataNode

1. **In a Hadoop cluster, what is the main role of the HDFS Client?**

* A) Managing and storing metadata
* B) Handling resource management tasks
* C) Interfacing with the HDFS to access data
* D) Executing MapReduce jobs
* **Answer:** C) Interfacing with the HDFS to access data

1. **What component in YARN manages the execution of containers for Hadoop jobs?**

* A) ResourceManager
* B) JobTracker
* C) NodeManager
* D) DataNode
* **Answer:** C) NodeManager

1. **Which of these components provides the actual physical storage of data in a Hadoop cluster?**

* A) ResourceManager
* B) JobTracker
* C) DataNode
* D) NameNode
* **Answer:** C) DataNode

1. **In Hadoop architecture, which of the following is responsible for resource allocation across different clusters?**

* A) YARN
* B) HDFS
* C) JobTracker
* D) ResourceManager
* **Answer:** A) YARN

1. **Which of the following Hadoop components allows for job execution on the cluster?**

* A) ResourceManager
* B) JobTracker
* C) NodeManager
* D) HBase
* **Answer:** B) JobTracker

1. **Which of the following is a major function of the ResourceManager in Hadoop’s YARN?**

* A) It monitors the health of nodes in the cluster.
* B) It schedules and allocates resources for job execution.
* C) It stores the metadata and coordinates file storage.
* D) It runs the actual computations on data blocks.
* **Answer:** B) It schedules and allocates resources for job execution.

1. **What is the function of the "TaskTracker" in Hadoop 1.x architecture?**

* A) It manages the file system namespace
* B) It executes the individual Map and Reduce tasks for a job
* C) It schedules the execution of jobs
* D) It stores data blocks in the HDFS
* **Answer:** B) It executes the individual Map and Reduce tasks for a job

1. **Which of the following is a core component of Hadoop that is responsible for parallel data processing?**

* A) HDFS
* B) MapReduce
* C) YARN
* D) Pig
* **Answer:** B) MapReduce

1. **What happens when a DataNode fails in Hadoop?**

* A) Data is lost permanently
* B) The NameNode replicates the data from other DataNodes
* C) The JobTracker takes over the responsibility of DataNode
* D) ResourceManager reallocates resources to the failed DataNode
* **Answer:** B) The NameNode replicates the data from other DataNodes

1. **Which Hadoop component is primarily responsible for handling job execution and task scheduling in the cluster?**

* A) JobTracker
* B) ResourceManager
* C) TaskTracker
* D) NodeManager
* **Answer:** A) JobTracker

### **Hard-Level Questions (26-40)**

1. **What happens in Hadoop when the ResourceManager runs out of resources for a new job?**

* A) The job is rejected and fails immediately.
* B) The job is placed in a queue until resources are available.
* C) The job is sent to the JobTracker for rescheduling.
* D) The job is sent to a backup ResourceManager.
* **Answer:** B) The job is placed in a queue until resources are available.

1. **Which of the following changes were introduced in Hadoop 2.x regarding the ResourceManager?**

* A) Multi-tenancy support
* B) Improved MapReduce performance
* C) Support for batch processing
* D) Removal of the JobTracker
* **Answer:** A) Multi-tenancy support

1. **In Hadoop's YARN architecture, which component is responsible for running jobs on slave nodes?**

* A) NodeManager
* B) ResourceManager
* C) JobTracker
* D) DataNode
* **Answer:** A) NodeManager

1. **Which of the following is the main limitation of Hadoop's original MapReduce framework?**

* A) Lack of real-time processing capabilities
* B) Inability to handle structured data
* C) Lack of parallel processing
* D) Inability to process large datasets
* **Answer:** A) Lack of real-time processing capabilities

1. **Which of the following statements is true about the HDFS architecture?**

* A) HDFS stores the metadata of files in DataNodes.
* B) Data is replicated in HDFS to ensure fault tolerance.
* C) All file system metadata is stored in DataNodes.
* D) Files are broken into fixed-size chunks, and each chunk is stored on a single node.
* **Answer:** B) Data is replicated in HDFS to ensure fault tolerance.

1. **How does YARN handle resource allocation in a multi-tenant environment?**

* A) By assigning resources to users in a round-robin fashion
* B) By allocating resources based on job priority
* C) By allowing each job to request the resources directly from DataNodes
* D) By assigning resources using the ResourceManager's scheduling algorithm
* **Answer:** D) By assigning resources using the ResourceManager's scheduling algorithm

1. **In a Hadoop 2.x cluster, what is the key advantage of using the YARN framework over the original MapReduce framework?**

* A) Better storage management
* B) Support for a variety of data processing frameworks
* C) Improved security mechanisms
* D) Support for only batch processing jobs
* **Answer:** B) Support for a variety of data processing frameworks

1. **Which of the following best describes Hadoop's "fault tolerance" feature?**

* A) Data is automatically backed up to a secondary cluster.
* B) Jobs are automatically rerun if a failure occurs.
* C) Data is replicated across multiple nodes to ensure availability.
* D) Nodes automatically recover from failures without manual intervention.
* **Answer:** C) Data is replicated across multiple nodes to ensure availability.

1. **What is the main benefit of Hadoop's "horizontal scalability"?**

* A) Easier to manage with fewer nodes
* B) Data can be processed without increasing the processing power
* C) New nodes can be added to the cluster without affecting system performance
* D) Improved processing time due to vertical scaling
* **Answer:** C) New nodes can be added to the cluster without affecting system performance

1. **In Hadoop, what is the maximum size of a block in HDFS by default?**

* A) 32MB
* B) 64MB
* C) 128MB
* D) 256MB
* **Answer:** C) 128MB

1. **Which of the following actions is NOT performed by the JobTracker in Hadoop 1.x?**

* A) Assigning tasks to TaskTrackers
* B) Monitoring the task progress
* C) Managing resource allocation
* D) Performing data storage management
* **Answer:** D) Performing data storage management

1. **How does Hadoop ensure that data is distributed evenly across multiple DataNodes?**

* A) By compressing data to make it smaller
* B) Through a hashing mechanism based on data block size
* C) By manually configuring each DataNode’s storage capacity
* D) By using a consistent hash function to distribute blocks
* **Answer:** B) Through a hashing mechanism based on data block size

1. **What is the purpose of the "Container" in Hadoop’s YARN framework?**

* A) It stores the metadata of the files being processed
* B) It is the environment where job execution occurs on each node
* C) It handles job scheduling across different nodes
* D) It replicates the data across multiple DataNodes
* **Answer:** B) It is the environment where job execution occurs on each node

1. **Which of the following is an example of a Hadoop-compatible file system other than HDFS?**

* A) HBase
* B) Amazon S3
* C) MapReduce
* D) HCatalog
* **Answer:** B) Amazon S3

1. **Which component in the Hadoop ecosystem is responsible for interacting with relational databases?**

* A) Hive
* B) Pig
* C) HBase
* D) Sqoop
* **Answer:** D) Sqoop

–##########################################################################

**Session 5:**

Setting up a Hadoop Cluster,

o Logging configuration

o Cluster specification,

o Cluster Setup and Installation,

o Common Hadoop Shell commands

o Clustering Monitoring

o Single and Multi-Node Cluster Setup on Virtual Machine,

o Hadoop Configuration, Security in Hadoop, Administering Hadoop,

o HDFS – Monitoring & Maintenance, Hadoop benchmarks

o Hadoop in the cloud.

1. **Which of the following is the primary requirement for setting up a Hadoop cluster?**
   * A) At least one computer with a single CPU
   * B) A set of virtual machines or physical machines
   * C) At least 100 GB of storage per node
   * D) A network connection with a speed of 10 Mbps or more
   * **Answer:** B) A set of virtual machines or physical machines
2. **Which of the following is used for logging configuration in Hadoop?**
   * A) log4j.properties
   * B) logging.xml
   * C) hadoop.log
   * D) logsettings.conf
   * **Answer:** A) log4j.properties
3. **What is the default configuration directory for Hadoop?**
   * A) /etc/hadoop/
   * B) /usr/local/hadoop/conf/
   * C) /hadoop/config/
   * D) /opt/hadoop/config/
   * **Answer:** B) /usr/local/hadoop/conf/
4. **Which of the following Hadoop components is used to monitor the health of the cluster?**
   * A) ResourceManager
   * B) HDFS Client
   * C) NameNode UI
   * D) JobTracker UI
   * **Answer:** C) NameNode UI
5. **Which of the following is a key feature of a multi-node Hadoop cluster?**
   * A) One machine stores all the data
   * B) Data is distributed across multiple machines
   * C) Only one DataNode is used
   * D) No need for MapReduce jobs
   * **Answer:** B) Data is distributed across multiple machines
6. **Which command is used to check the status of HDFS in Hadoop?**
   * A) hadoop status
   * B) hdfs dfsadmin -report
   * C) hadoop dfs status
   * D) hdfs status
   * **Answer:** B) hdfs dfsadmin -report
7. **Which file is used for setting the cluster specifications in Hadoop?**
   * A) hadoop-env.sh
   * B) core-site.xml
   * C) hdfs-site.xml
   * D) mapred-site.xml
   * **Answer:** B) core-site.xml
8. **Which of the following is a common Hadoop shell command to list the contents of a directory in HDFS?**
   * A) hadoop fs -ls
   * B) hdfs dfs -ls
   * C) hadoop ls
   * D) fs -list
   * **Answer:** B) hdfs dfs -ls
9. **When setting up a Hadoop cluster, which of the following is important for performance optimization?**
   * A) Installing Hadoop on a single node
   * B) Increasing the number of DataNodes
   * C) Installing multiple ResourceManagers
   * D) Increasing the size of the NameNode
   * **Answer:** B) Increasing the number of DataNodes
10. **In a multi-node Hadoop cluster, what is the role of the secondary NameNode?**

* A) Backup NameNode for fault tolerance
* B) It acts as a backup for ResourceManager
* C) It runs MapReduce jobs
* D) It manages the user access permissions
* **Answer:** A) Backup NameNode for fault tolerance

### **Intermediate-Level Questions (11-25)**

1. **Which of the following Hadoop shell commands is used to move data between local filesystem and HDFS?**

* A) hdfs dfs -get
* B) hdfs dfs -put
* C) hadoop fs -move
* D) Both A and B
* **Answer:** D) Both A and B

1. **In Hadoop, what is the role of the JobTracker in the cluster setup?**

* A) To monitor data replication
* B) To schedule jobs and distribute tasks across TaskTrackers
* C) To allocate resources to DataNodes
* D) To store metadata information
* **Answer:** B) To schedule jobs and distribute tasks across TaskTrackers

1. **Which of the following is the first step in setting up a Hadoop cluster?**

* A) Install MapReduce
* B) Configure HDFS settings
* C) Install Java and Hadoop binaries
* D) Set up network connectivity
* **Answer:** C) Install Java and Hadoop binaries

1. **What is the main purpose of the hdfs dfsadmin -safemode enter command?**

* A) To activate HDFS safe mode to prevent data corruption
* B) To perform a cluster-wide health check
* C) To forcefully remove files from HDFS
* D) To check the storage health of the DataNodes
* **Answer:** A) To activate HDFS safe mode to prevent data corruption

1. **Which of the following is used to monitor a Hadoop cluster’s resource usage in YARN?**

* A) YARN ResourceManager UI
* B) JobTracker UI
* C) HDFS UI
* D) NameNode UI
* **Answer:** A) YARN ResourceManager UI

1. **Which of the following is required to install Hadoop on a virtual machine?**

* A) At least two virtual machines
* B) A cluster of physical machines
* C) A Hadoop-compatible file system
* D) A pre-configured cloud service
* **Answer:** A) At least two virtual machines

1. **Which Hadoop configuration file is responsible for configuring the HDFS block size?**

* A) core-site.xml
* B) mapred-site.xml
* C) hdfs-site.xml
* D) hadoop-env.sh
* **Answer:** C) hdfs-site.xml

1. **Which of the following describes Hadoop security features?**

* A) Kerberos authentication
* B) Data encryption at rest and in transit
* C) User authorization through Hadoop’s access control
* D) All of the above
* **Answer:** D) All of the above

1. **How does Hadoop handle large data processing efficiently?**

* A) By dividing data into blocks and processing them in parallel
* B) By storing data in a single node
* C) By using only sequential processing
* D) By compressing the data before processing
* **Answer:** A) By dividing data into blocks and processing them in parallel

1. **What command is used to start the Hadoop daemons (NameNode, DataNode, ResourceManager, etc.)?**

* A) start-dfs.sh
* B) start-hadoop.sh
* C) hadoop start all
* D) hadoop start cluster
* **Answer:** A) start-dfs.sh

1. **Which of the following commands is used to stop the Hadoop daemons?**

* A) stop-dfs.sh
* B) hadoop stop cluster
* C) stop-all.sh
* D) stop-hadoop.sh
* **Answer:** A) stop-dfs.sh

1. **Which of the following is a common Hadoop benchmark tool used to test the performance of the cluster?**

* A) TeraSort
* B) Hadoop-bench
* C) MapReduce-bench
* D) HDFS-bench
* **Answer:** A) TeraSort

1. **What is the main purpose of using a multi-node cluster over a single-node cluster in Hadoop?**

* A) To improve fault tolerance and scalability
* B) To reduce network latency
* C) To simplify cluster setup
* D) To process data more quickly
* **Answer:** A) To improve fault tolerance and scalability

1. **In a Hadoop cluster, what is the role of the HDFS balancer?**

* A) To balance data replication across DataNodes
* B) To balance storage capacity across DataNodes
* C) To manage job scheduling across nodes
* D) To encrypt data at rest
* **Answer:** B) To balance storage capacity across DataNodes

### **Hard-Level Questions (26-40)**

1. **Which configuration file should you modify to specify the port number for the NameNode in Hadoop?**

* A) mapred-site.xml
* B) hdfs-site.xml
* C) core-site.xml
* D) hadoop-env.sh
* **Answer:** B) hdfs-site.xml

1. **What is the key advantage of setting up a Hadoop cluster in the cloud?**

* A) Reduced cost of hardware
* B) Scalability based on demand
* C) Easier cluster setup and management
* D) All of the above
* **Answer:** D) All of the above

1. **In Hadoop, which of the following tools can be used to import data from relational databases?**

* A) Sqoop
* B) Flume
* C) Oozie
* D) Hive
* **Answer:** A) Sqoop

1. **What is the primary role of the ResourceManager in a YARN-enabled Hadoop cluster?**

* A) To allocate resources to applications
* B) To manage the HDFS
* C) To schedule jobs in MapReduce
* D) To manage the distributed cache
* **Answer:** A) To allocate resources to applications

1. **Which of the following is true about Hadoop in the cloud?**

* A) Hadoop cannot be run in cloud environments
* B) Cloud-based Hadoop clusters offer automatic scaling
* C) Hadoop in the cloud only works with AWS
* D) Cloud-based clusters are limited to small-scale data processing
* **Answer:** B) Cloud-based Hadoop clusters offer automatic scaling

1. **In Hadoop security, what is the purpose of Kerberos authentication?**

* A) To encrypt data in transit
* B) To authenticate users and nodes in the cluster
* C) To prevent unauthorized data replication
* D) To manage the ResourceManager access control
* **Answer:** B) To authenticate users and nodes in the cluster

1. **What is the benefit of using a dedicated HDFS cluster in the cloud for processing large datasets?**

* A) Unlimited storage capacity with high availability
* B) Reduced overhead due to resource isolation
* C) Lower cost compared to on-premise clusters
* D) All of the above

1. **Which of the following is a critical part of administering a Hadoop cluster?**

* A) Configuring YARN memory settings
* B) Managing Hadoop users and permissions
* C) Monitoring job and resource performance
* D) All of the above
* **Answer:** D) All of the above

1. **How does the HDFS balancer determine which DataNodes to move data between?**

* A) Based on the node’s memory capacity
* B) Based on the storage space used across nodes
* C) Based on CPU usage of the DataNodes
* D) Based on network speed
* **Answer:** B) Based on the storage space used across nodes

1. **When running a benchmark using TeraSort in Hadoop, which aspect of the cluster is primarily tested?**

* A) HDFS throughput
* B) YARN resource allocation
* C) Data processing speed
* D) Network latency
* **Answer:** C) Data processing speed

1. **Which of the following is the main role of the ResourceManager in a Hadoop cluster running YARN?**

* A) Monitor cluster health
* B) Run MapReduce jobs
* C) Manage and schedule resources
* D) Monitor HDFS health
* **Answer:** C) Manage and schedule resources

1. **Which of the following is true for Hadoop benchmarks like TeraSort and Sort?**

* A) They test the efficiency of distributed file systems
* B) They test the memory capacity of individual nodes
* C) They assess the performance of the Hadoop MapReduce framework
* D) They only measure CPU usage
* **Answer:** C) They assess the performance of the Hadoop MapReduce framework

1. **What is a key consideration when configuring Hadoop security?**

* A) Ensuring job replication is enabled
* B) Configuring data encryption settings
* C) Using secure node-to-node communication
* D) Both B and C
* **Answer:** D) Both B and C

1. **What does Hadoop’s "fault tolerance" mean?**

* A) Data is automatically backed up to another system
* B) The system recovers automatically when a node fails
* C) All tasks are automatically rerun if they fail
* D) Data is replicated to ensure system integrity
* **Answer:** D) Data is replicated to ensure system integrity

1. **Which is an important factor when setting up Hadoop in a cloud environment?**

* A) Ensuring that on-premise hardware is compatible
* B) Configuring cloud storage policies and access rights
* C) Only using local clusters
* D) Avoiding the use of external storage systems
* **Answer:** B) Configuring cloud storage policies and access rights

##################################################################################

**Session6**

**Hadoop Distributed File System (HDFS)**

o Distributed File System,

o What is HDFS,

o Major goals of HDFS Design

o Where does HDFS fit in,

o Core components of HDFS,

o Hadoop Server Roles: Name Node, Secondary Name Node, and Data Node

### **Easy-Level Questions (1-10)**

1. **What does HDFS stand for in the context of Hadoop?**
   * A) Hadoop Data Flow System
   * B) Hadoop Distributed File System
   * C) High Data File System
   * D) Hadoop File Storage
   * **Answer:** B) Hadoop Distributed File System
2. **Which of the following best describes the main function of HDFS?**
   * A) Distributed processing of large datasets
   * B) Storage of large datasets across multiple machines
   * C) Management of job execution across a cluster
   * D) Data replication across a single machine
   * **Answer:** B) Storage of large datasets across multiple machines
3. **What is the primary goal of HDFS?**
   * A) Efficient file processing
   * B) High availability and scalability
   * C) Real-time data processing
   * D) Data compression
   * **Answer:** B) High availability and scalability
4. **In Hadoop, where does HDFS fit into the overall ecosystem?**
   * A) As the main tool for processing big data
   * B) As the storage layer for large datasets
   * C) As the job scheduler for MapReduce tasks
   * D) As the tool for data replication
   * **Answer:** B) As the storage layer for large datasets
5. **Which of the following is the primary function of the DataNode in HDFS?**
   * A) To manage the job execution
   * B) To store data in the form of blocks
   * C) To replicate data blocks for fault tolerance
   * D) To provide access to the NameNode
   * **Answer:** B) To store data in the form of blocks
6. **Which HDFS component is responsible for storing metadata about the file system?**
   * A) Secondary NameNode
   * B) DataNode
   * C) NameNode
   * D) ResourceManager
   * **Answer:** C) NameNode
7. **What is the role of the Secondary NameNode in HDFS?**
   * A) To manage the cluster’s jobs and resources
   * B) To act as a backup for the NameNode’s metadata
   * C) To process and store user data
   * D) To monitor the status of the DataNodes
   * **Answer:** B) To act as a backup for the NameNode’s metadata
8. **Which of the following is a key feature of HDFS?**
   * A) Data is divided into large blocks and distributed across multiple nodes
   * B) Files are stored in a single location for better accessibility
   * C) Data is processed in real-time by MapReduce
   * D) HDFS stores data in uncompressed format
   * **Answer:** A) Data is divided into large blocks and distributed across multiple nodes
9. **In HDFS, what is the typical size of a block?**
   * A) 1 KB
   * B) 64 MB
   * C) 128 MB
   * D) 256 MB
   * **Answer:** C) 128 MB
10. **Which of the following is a critical advantage of using HDFS for large data storage?**

* A) Centralized storage on a single node
* B) Automatic backup of all files
* C) High scalability and fault tolerance
* D) Real-time data analysis
* **Answer:** C) High scalability and fault tolerance

### **Intermediate-Level Questions (11-25)**

1. **What happens when a DataNode in HDFS fails?**

* A) Data is automatically restored from the Secondary NameNode
* B) The NameNode replicates data from other DataNodes to replace the lost data
* C) The cluster crashes until the DataNode is fixed
* D) The DataNode restarts automatically with the same data
* **Answer:** B) The NameNode replicates data from other DataNodes to replace the lost data

1. **Which of the following HDFS components is responsible for managing the file system namespace?**

* A) DataNode
* B) JobTracker
* C) NameNode
* D) ResourceManager
* **Answer:** C) NameNode

1. **Which of the following best describes the role of HDFS in the Hadoop ecosystem?**

* A) HDFS handles computation across distributed systems
* B) HDFS provides an interface for programming MapReduce jobs
* C) HDFS stores large datasets across a distributed network of computers
* D) HDFS processes data stored on a single machine
* **Answer:** C) HDFS stores large datasets across a distributed network of computers

1. **Which of the following is the default replication factor for data stored in HDFS?**

* A) 1
* B) 2
* C) 3
* D) 5
* **Answer:** C) 3

1. **What does HDFS use to ensure fault tolerance?**

* A) Data compression
* B) Data replication across multiple nodes
* C) Single-node backups
* D) Data encryption
* **Answer:** B) Data replication across multiple nodes

1. **Which of the following is NOT a core component of HDFS?**

* A) NameNode
* B) DataNode
* C) ResourceManager
* D) Secondary NameNode
* **Answer:** C) ResourceManager

1. **Which of the following components interacts with HDFS to schedule jobs and handle job execution?**

* A) ResourceManager
* B) JobTracker
* C) DataNode
* D) YARN
* **Answer:** A) ResourceManager

1. **How does HDFS achieve high availability?**

* A) By using multiple copies of data blocks
* B) By using distributed nodes
* C) By storing data in real-time
* D) By caching data on the local filesystem
* **Answer:** A) By using multiple copies of data blocks

1. **What is the main disadvantage of using small file sizes in HDFS?**

* A) Poor network bandwidth usage
* B) Inability to perform parallel processing
* C) Overhead in managing a large number of files
* D) Slow replication times
* **Answer:** C) Overhead in managing a large number of files

1. **Which of the following is true about the Secondary NameNode?**

* A) It is a backup of the DataNode
* B) It stores metadata for fast lookup
* C) It periodically checkpoints the NameNode metadata
* D) It performs job execution tasks
* **Answer:** C) It periodically checkpoints the NameNode metadata

1. **Which command in HDFS is used to copy files from the local file system to HDFS?**

* A) hdfs dfs -put
* B) hdfs dfs -get
* C) hadoop fs -copyToLocal
* D) hdfs dfs -cp
* **Answer:** A) hdfs dfs -put

1. **In HDFS, which component is responsible for managing block placement across DataNodes?**

* A) NameNode
* B) DataNode
* C) Secondary NameNode
* D) ResourceManager
* **Answer:** A) NameNode

1. **What is the function of the fsimage in HDFS?**

* A) It stores actual data
* B) It stores the file system metadata
* C) It is used for data replication
* D) It stores HDFS block information
* **Answer:** B) It stores the file system metadata

1. **What happens when the replication factor of a block in HDFS is reduced?**

* A) Data becomes highly available
* B) Data is compressed
* C) The block may become under-replicated
* D) Data becomes more secure
* **Answer:** C) The block may become under-replicated

1. **Which of the following commands is used to display the HDFS file system status?**

* A) hdfs dfs -status
* B) hdfs dfsadmin -report
* C) hadoop fs -status
* D) hdfs dfs -health
* **Answer:** B) hdfs dfsadmin -report

### **Hard-Level Questions (26-40)**

1. **What is the role of the Block Scanner in HDFS?**

* A) To ensure that all blocks are replicated correctly
* B) To scan and repair corrupt blocks
* C) To compress HDFS data blocks
* D) To manage block size allocation
* **Answer:** B) To scan and repair corrupt blocks

1. **In HDFS, what is the function of the Data Transfer Protocol?**

* A) It ensures that DataNodes communicate with the NameNode
* B) It enables DataNodes to transfer block data between them
* C) It manages data encryption
* D) It helps in data compression
* **Answer:** B) It enables DataNodes to transfer block data between them

1. **How does HDFS achieve write-once-read-many (WORM) semantics?**

* A) By allowing multiple write operations to the same block
* B) By preventing modification of data once written
* C) By storing data in compressed format
* D) By using version control for data blocks
* **Answer:** B) By preventing modification of data once written

1. **In HDFS, what happens when a DataNode fails and there are insufficient replicas?**

* A) Data is lost permanently
* B) Data is replicated from other healthy DataNodes
* C) The NameNode tries to recover the data
* D) The entire cluster crashes
* **Answer:** B) Data is replicated from other healthy DataNodes

1. **Which of the following statements is true about block placement policy in HDFS?**

* A) Blocks are placed on DataNodes in a round-robin fashion
* B) Data is placed on nodes with the least disk usage
* C) HDFS tries to place replicas on different racks for fault tolerance
* D) All blocks are placed on the same node to ensure faster access
* **Answer:** C) HDFS tries to place replicas on different racks for fault tolerance

1. **Which of the following components ensures the consistency of HDFS?**

* A) DataNode
* B) NameNode
* C) Secondary NameNode
* D) ResourceManager
* **Answer:** B) NameNode

1. **What does the HDFS balancer tool do?**

* A) Balances the data across all DataNodes based on disk space utilization
* B) Balances job execution across nodes
* C) Replicates data from one node to another
* D) Compresses data blocks across DataNodes
* **Answer:** A) Balances the data across all DataNodes based on disk space utilization

1. **How does HDFS handle block replication when a node is added to the cluster?**

* A) HDFS automatically removes blocks from the new node
* B) Blocks from under-replicated nodes are moved to the new node
* C) The block size is reduced for new nodes
* D) No changes are made to the block replication policy
* **Answer:** B) Blocks from under-replicated nodes are moved to the new node

1. **What is the Block Report in HDFS?**

* A) A list of blocks being transferred between DataNodes
* B) A log of block replication failures
* C) A periodic report generated by DataNodes to inform the NameNode about the status of blocks
* D) A summary of the blocks stored in HDFS
* **Answer:** C) A periodic report generated by DataNodes to inform the NameNode about the status of blocks

1. **Which of the following is true about HDFS data block sizes?**

* A) Block size is fixed for all types of data
* B) Block size must be smaller than 1 MB for better performance
* C) Block sizes can be configured depending on the type of data and application requirements
* D) Block size is irrelevant in HDFS
* **Answer:** C) Block sizes can be configured depending on the type of data and application requirements

1. **What is the role of the File System Namespace in HDFS?**

* A) To store metadata for blocks
* B) To store a hierarchy of files and directories
* C) To keep track of DataNode storage health
* D) To manage block replication policies
* **Answer:** B) To store a hierarchy of files and directories

1. **Which command is used to check the health of an HDFS cluster?**

* A) hdfs dfs -status
* B) hdfs dfsadmin -health
* C) hadoop fs -health
* D) hdfs dfsadmin -report
* **Answer:** B) hdfs dfsadmin -health

1. **How does the HDFS NameNode handle consistency when multiple users are accessing the system?**

* A) It ensures data consistency through locking mechanisms
* B) It uses version control for files
* C) It does not manage consistency; it is handled by DataNodes
* D) It maintains consistency through distributed file metadata
* **Answer:** D) It maintains consistency through distributed file metadata

1. **Which of the following is a disadvantage of using HDFS for storing large files?**

* A) The system is highly prone to failure
* B) Data replication consumes a lot of storage
* C) Only small files can be stored efficiently
* D) It is slower than other distributed systems
* **Answer:** B) Data replication consumes a lot of storage

#########################################################################

Session 7.

**HDFS Architecture**

o HDFS Architecture,

o Scaling and Rebalancing,

o Big Deal about HDFS,

o Replication,

o Rack Awareness,

o Data Pipelining,

o Node Failure Management.

o HDFS NameNode High Availability

o Components and daemon of an HDFS HA-Quorum cluster

o HDFS Federation use case

o Kerberos: Role of HDFS security

### **Easy-Level Questions (1-10)**

1. **What is the primary role of HDFS in the Hadoop ecosystem?**
   * A) Job execution
   * B) Data storage and management
   * C) Data processing
   * D) Resource management
   * **Answer:** B) Data storage and management
2. **What is the basic unit of storage in HDFS?**
   * A) File
   * B) DataNode
   * C) Block
   * D) Directory
   * **Answer:** C) Block
3. **What is the typical block size used in HDFS?**
   * A) 1 MB
   * B) 64 MB
   * C) 128 MB
   * D) 256 KB
   * **Answer:** C) 128 MB
4. **In HDFS, what does replication refer to?**
   * A) Storing the same data on multiple DataNodes for fault tolerance
   * B) Storing data on a single node for faster access
   * C) Compressing data for storage efficiency
   * D) Encrypting data before storage
   * **Answer:** A) Storing the same data on multiple DataNodes for fault tolerance
5. **What is the goal of HDFS scaling?**
   * A) To improve job execution speed
   * B) To increase storage capacity by adding nodes
   * C) To replicate data for redundancy
   * D) To improve real-time processing capabilities
   * **Answer:** B) To increase storage capacity by adding nodes
6. **Which of the following HDFS features helps manage large amounts of data across nodes?**
   * A) Rack awareness
   * B) File compression
   * C) Real-time data processing
   * D) Data pipelining
   * **Answer:** A) Rack awareness
7. **What is the role of rack awareness in HDFS?**
   * A) It determines which DataNodes are in the same physical rack
   * B) It determines the file compression strategy
   * C) It determines the block size for each file
   * D) It determines the number of file replications
   * **Answer:** A) It determines which DataNodes are in the same physical rack
8. **What is data pipelining in HDFS?**
   * A) The process of transferring data between nodes in parallel
   * B) The process of copying data blocks across the network
   * C) The mechanism by which data is read and written in sequential blocks
   * D) The system of encrypting data before storage
   * **Answer:** C) The mechanism by which data is read and written in sequential blocks
9. **Which of the following is true about node failure management in HDFS?**
   * A) Data is lost if a DataNode fails
   * B) The NameNode automatically restores the data on failed nodes
   * C) The data is automatically replicated to other nodes when a DataNode fails
   * D) Data recovery requires manual intervention by administrators
   * **Answer:** C) The data is automatically replicated to other nodes when a DataNode fails
10. **What is the purpose of NameNode High Availability (HA) in HDFS?**

* A) To store all the data in a redundant manner
* B) To ensure the NameNode remains available in case of failure
* C) To replicate data across multiple nodes
* D) To improve job execution speed
* **Answer:** B) To ensure the NameNode remains available in case of failure

### **Intermediate-Level Questions (11-25)**

1. **Which of the following is true regarding HDFS federation?**

* A) It allows for the storage of data on a single NameNode
* B) It enables horizontal scalability by using multiple NameNodes
* C) It limits the number of DataNodes to a single cluster
* D) It provides automatic backup of data blocks
* **Answer:** B) It enables horizontal scalability by using multiple NameNodes

1. **What does HDFS quorum refer to in a high-availability setup?**

* A) The number of DataNodes in the cluster
* B) The minimum number of replicas required for a block
* C) The consensus mechanism used to maintain cluster integrity in HA setups
* D) The process of distributing data across racks
* **Answer:** C) The consensus mechanism used to maintain cluster integrity in HA setups

1. **In a high-availability HDFS setup, what is the role of the JournalNode?**

* A) It replicates data blocks across nodes
* B) It maintains transaction logs for the NameNode
* C) It monitors the health of DataNodes
* D) It executes MapReduce jobs
* **Answer:** B) It maintains transaction logs for the NameNode

1. **Which of the following components is NOT part of the HDFS High Availability architecture?**

* A) Active NameNode
* B) Standby NameNode
* C) JournalNode
* D) ResourceManager
* **Answer:** D) ResourceManager

1. **What happens when the Active NameNode fails in an HDFS High Availability setup?**

* A) Data is lost until the system is restored
* B) The standby NameNode automatically takes over
* C) The cluster stops accepting new data
* D) The JournalNode replicates data from another cluster
* **Answer:** B) The standby NameNode automatically takes over

1. **Which of the following is a key advantage of HDFS Federation?**

* A) Simplified data replication process
* B) Improved data integrity checks
* C) Scalability by adding more NameNodes
* D) Faster job execution times
* **Answer:** C) Scalability by adding more NameNodes

1. **In HDFS, Kerberos authentication is used for what purpose?**

* A) To encrypt data before storing it in HDFS
* B) To authenticate users and nodes accessing the HDFS cluster
* C) To ensure high availability of the NameNode
* D) To manage job execution in the cluster
* **Answer:** B) To authenticate users and nodes accessing the HDFS cluster

1. **Which of the following is a benefit of using rack awareness in HDFS?**

* A) Optimizing block replication across racks for fault tolerance
* B) Ensuring all data is stored in a single location for easy access
* C) Enabling real-time data processing across racks
* D) Compressing data on the same rack to save space
* **Answer:** A) Optimizing block replication across racks for fault tolerance

1. **Which command is used to check the status of an HDFS cluster?**

* A) hdfs dfsadmin -report
* B) hdfs dfs -status
* C) hadoop fs -checkstatus
* D) hdfs fs -health
* **Answer:** A) hdfs dfsadmin -report

1. **What is HDFS rebalancing used for?**

* A) Balancing the data load across different nodes based on usage
* B) Re-distributing data blocks across racks
* C) Moving data from underutilized DataNodes to more active ones
* D) Both A and C
* **Answer:** D) Both A and C

1. **What is the effect of NameNode high availability on the failover process?**

* A) It speeds up the replication of data
* B) It ensures the active NameNode is always available
* C) It guarantees that all data is instantly replicated
* D) It helps in compressing data for quicker access
* **Answer:** B) It ensures the active NameNode is always available

1. **What is the role of the Standby NameNode in an HA setup?**

* A) It is the main server for handling all file system operations
* B) It maintains a copy of the active NameNode’s state and becomes active during failover
* C) It stores block replicas in case of node failure
* D) It handles data replication across the entire cluster
* **Answer:** B) It maintains a copy of the active NameNode’s state and becomes active during failover

1. **What happens when a DataNode fails in HDFS?**

* A) The entire HDFS cluster fails
* B) The data is automatically replicated to other healthy DataNodes
* C) The system stops accepting new data until recovery
* D) Data stored on the failed DataNode is lost forever
* **Answer:** B) The data is automatically replicated to other healthy DataNodes

1. **Which of the following is true about HDFS Federation?**

* A) It can have multiple NameNodes for different namespaces
* B) It limits the number of blocks stored across the DataNodes
* C) It provides fault tolerance by replicating data within a single namespace
* D) It only allows one NameNode to be active at a time
* **Answer:** A) It can have multiple NameNodes for different namespaces

1. **In HDFS, which of the following is a critical consideration for configuring Kerberos?**

* A) Ensuring that the keytab files are securely stored
* B) Ensuring that data is stored in a compressed format
* C) Ensuring that data is split into smaller blocks
* D) Ensuring that the system has an active standby NameNode
* **Answer:** A) Ensuring that the keytab files are securely stored

### **Advanced-Level Questions (26-50)**

1. **Which of the following is a characteristic of HDFS Federation?**

* A) It stores all metadata in a single NameNode
* B) It allows multiple independent NameNodes to manage different namespaces
* C) It reduces the need for replication of data blocks
* D) It stores data across only one rack for performance reasons
* **Answer:** B) It allows multiple independent NameNodes to manage different namespaces

1. **In an HDFS cluster with multiple NameNodes, which component manages the block management across namespaces?**

* A) JournalNode
* B) DataNode
* C) ResourceManager
* D) Block Manager
* **Answer:** A) JournalNode

1. **Which of the following describes node failure management in HDFS?**

* A) Data from a failed node is lost and cannot be recovered
* B) DataNodes replicate data automatically to other healthy nodes when a failure occurs
* C) NameNode automatically reassigns all blocks to new nodes when a failure occurs
* D) Data replication occurs only during cluster shutdown
* **Answer:** B) DataNodes replicate data automatically to other healthy nodes when a failure occurs

1. **What is the primary purpose of rack awareness in HDFS?**

* A) To optimize data retrieval times across different racks
* B) To place replicas of data blocks in a way that reduces the likelihood of losing data during rack failures
* C) To ensure that data is replicated on nodes within the same rack for faster access
* D) To enable automatic rebalancing of data during rack failures
* **Answer:** B) To place replicas of data blocks in a way that reduces the likelihood of losing data during rack failures

1. **Which of the following is a key advantage of HDFS data pipelining?**

* A) It allows blocks to be written sequentially from multiple clients
* B) It minimizes network congestion by transferring blocks in parallel across DataNodes
* C) It optimizes data block size by using multiple compression algorithms
* D) It allows data to be read and written at the same time across multiple nodes
* **Answer:** A) It allows blocks to be written sequentially from multiple clients

1. **What is the role of the Secondary NameNode in HDFS?**

* A) It maintains a checkpoint of the HDFS metadata by periodically merging the namespace image with the edit log
* B) It takes over the role of the primary NameNode during failover
* C) It manages the replication of data blocks across DataNodes
* D) It stores a backup of the cluster’s data blocks
* **Answer:** A) It maintains a checkpoint of the HDFS metadata by periodically merging the namespace image with the edit log

1. **In an HDFS high availability (HA) setup, how is the failover process triggered?**

* A) Automatically by the NameNode when it detects data corruption
* B) Manually by the cluster administrator during maintenance
* C) Automatically through the Zookeeper service, which manages the active and standby states
* D) It occurs only when the Secondary NameNode fails
* **Answer:** C) Automatically through the Zookeeper service, which manages the active and standby states

1. **What is the role of Zookeeper in HDFS high availability?**

* A) It manages block replication across multiple DataNodes
* B) It keeps track of the current active and standby NameNodes
* C) It monitors data integrity during writes to HDFS
* D) It helps distribute data blocks across different racks
* **Answer:** B) It keeps track of the current active and standby NameNodes

1. **Which of the following best describes HDFS rebalancing?**

* A) It redistributes blocks to ensure an even storage load across DataNodes
* B) It compresses blocks to save space on DataNodes
* C) It adjusts the replication factor dynamically for each file
* D) It removes blocks from underutilized DataNodes to free up space
* **Answer:** A) It redistributes blocks to ensure an even storage load across DataNodes

1. **Which tool in HDFS can be used to perform rebalancing operations?**

* A) hdfs dfsadmin -reassign
* B) hdfs balancer
* C) hdfs fsck
* D) hdfs dfsadmin -rebalance
* **Answer:** B) hdfs balancer

1. \*\*How does **HDFS handle node failure in terms of block replication?**

* A) The data becomes unavailable until the node is restored
* B) HDFS automatically copies the lost block from other replicas to maintain the replication factor
* C) The failed node is removed, and data is permanently lost
* D) HDFS stops all writes until the failed node is brought back online
* **Answer:** B) HDFS automatically copies the lost block from other replicas to maintain the replication factor

1. **In the context of Kerberos authentication, what does the Keytab file contain?**

* A) The data block hashes used for encryption
* B) The access control policies for DataNodes
* C) The encrypted credentials used by clients and nodes for authentication
* D) A list of authorized users and their roles
* **Answer:** C) The encrypted credentials used by clients and nodes for authentication

1. **What is the key advantage of using Kerberos authentication in HDFS?**

* A) It reduces the complexity of managing user permissions
* B) It ensures secure communication between clients, DataNodes, and the NameNode
* C) It allows for anonymous access to data stored in HDFS
* D) It eliminates the need for replication of data blocks
* **Answer:** B) It ensures secure communication between clients, DataNodes, and the NameNode

1. **How does HDFS Federation improve scalability?**

* A) By consolidating multiple NameNodes into a single namespace
* B) By reducing the replication factor for blocks across different namespaces
* C) By allowing the addition of more NameNodes, thus distributing the metadata load
* D) By automatically balancing data across racks
* **Answer:** C) By allowing the addition of more NameNodes, thus distributing the metadata load

1. **Which component is responsible for tracking the status of blocks and their replicas in an HDFS cluster?**

* A) ResourceManager
* B) DataNode
* C) NameNode
* D) JobTracker
* **Answer:** C) NameNode

1. **What is the main advantage of using HDFS High Availability in a production environment?**

* A) It allows multiple clients to read from the same block simultaneously
* B) It ensures that the system remains available even if the active NameNode fails
* C) It reduces the replication factor of blocks to save storage space
* D) It increases the speed of job execution in MapReduce
* **Answer:** B) It ensures that the system remains available even if the active NameNode fails

1. **Which of the following is a feature of HDFS NameNode High Availability?**

* A) The NameNode is replicated across multiple racks
* B) The secondary NameNode is responsible for block management
* C) The Zookeeper service helps manage the active and standby NameNode states
* D) There is no impact on system performance during a failover
* **Answer:** C) The Zookeeper service helps manage the active and standby NameNode states

1. **What is the main role of Zookeeper in HDFS NameNode High Availability?**

* A) To handle block replication
* B) To monitor and switch between active and standby NameNodes
* C) To store metadata for HDFS files
* D) To ensure data compression during writes
* **Answer:** B) To monitor and switch between active and standby NameNodes

1. **Which of the following is true about rack awareness in an HDFS setup?**

* A) It helps to optimize the storage of metadata in racks
* B) It ensures that data blocks are replicated on different racks for fault tolerance
* C) It allows for automatic storage of data on the same rack as the client
* D) It eliminates the need for multiple replicas of data
* **Answer:** B) It ensures that data blocks are replicated on different racks for fault tolerance

1. **Which of the following commands will show the health of an HDFS cluster?**

* A) hdfs dfsadmin -report
* B) hdfs dfs -health
* C) hdfs fsck
* D) hdfs checkstatus
* **Answer:** A) hdfs dfsadmin -report

1. **What does HDFS pipelining facilitate?**

* A) Parallel block writes to improve throughput
* B) Block replication between racks
* C) Sequential block writes across multiple DataNodes
* D) Data compression during storage
* **Answer:** C) Sequential block writes across multiple DataNodes

1. **Which of the following is true about HDFS block replication?**

* A) It is done to improve the read/write performance of blocks
* B) It reduces the storage space usage in the cluster
* C) It ensures fault tolerance by creating copies of data on different nodes
* D) It limits the size of data blocks stored in a node
* **Answer:** C) It ensures fault tolerance by creating copies of data on different nodes

1. **What action is triggered when a DataNode detects a failure in HDFS?**

* A) It marks the block as unavailable and deletes it
* B) It replicates the block to another healthy DataNode
* C) It sends an alert to the NameNode
* D) It immediately starts rebalancing the data
* **Answer:** B) It replicates the block to another healthy DataNode

1. **What happens if the replication factor is reduced in HDFS?**

* A) Data becomes less secure
* B) There is an increase in disk usage
* C) It saves space by storing fewer replicas of the data
* D) The system becomes less fault-tolerant
* **Answer:** C) It saves space by storing fewer replicas of the data

#################################################################################

**Session 8**

**HDFS Data Storage Process**

**o HDFS Data storage process,**

**o Anatomy of writing and reading file in HDFS,**

**o HDFS user and admin commands,**

**o HDFS Web Interface.**

### **HDFS Data Storage Process MCQs**

1. **Which of the following is the first step in the HDFS data storage process when a client writes a file?**
   * A) The file is split into fixed-size blocks and stored on DataNodes
   * B) The NameNode assigns the file’s blocks to DataNodes for storage
   * C) The file is stored in the HDFS Web Interface
   * D) The DataNode replicates the file to multiple servers
   * **Answer:** B) The NameNode assigns the file’s blocks to DataNodes for storage
2. **In HDFS, which component is responsible for maintaining metadata about the stored files, including file names, block locations, and permissions?**
   * A) DataNode
   * B) ResourceManager
   * C) NameNode
   * D) JobTracker
   * **Answer:** C) NameNode
3. **In the context of HDFS file writing, which action is taken when a client sends data to the NameNode?**
   * A) The NameNode replicates the data on multiple DataNodes
   * B) The NameNode selects DataNodes and sends block locations to the client
   * C) The NameNode stores the data directly on its disk
   * D) The DataNodes return data to the client immediately
   * **Answer:** B) The NameNode selects DataNodes and sends block locations to the client
4. **How does HDFS handle block replication when a DataNode fails during the writing process?**
   * A) The write operation fails, and data is not stored
   * B) The data is replicated on other DataNodes to meet the replication factor
   * C) The NameNode marks the block as unreplicated until the DataNode recovers
   * D) The data is written to an alternate NameNode
   * **Answer:** B) The data is replicated on other DataNodes to meet the replication factor
5. **What is the typical anatomy of writing a file to HDFS?**
   * A) The client communicates with the DataNode and writes the file directly
   * B) The client communicates with the NameNode, which provides DataNode locations; data is then written to the DataNodes in blocks
   * C) The file is first compressed by the client before writing to the DataNode
   * D) The NameNode writes the file directly to HDFS, bypassing DataNodes
   * **Answer:** B) The client communicates with the NameNode, which provides DataNode locations; data is then written to the DataNodes in blocks
6. **Which of the following describes HDFS reading file anatomy?**
   * A) The client first reads the file from the NameNode, which stores the file content
   * B) The client communicates with the NameNode, retrieves the block locations, and reads the blocks directly from the DataNodes
   * C) The client sends data to the DataNode, which sends it back to the client
   * D) The client writes data to HDFS, and the NameNode retrieves it for access
   * **Answer:** B) The client communicates with the NameNode, retrieves the block locations, and reads the blocks directly from the DataNodes
7. **In the HDFS read process, how does the client handle block retrieval from the DataNodes?**
   * A) The client reads the file sequentially from the first DataNode
   * B) The client reads all blocks from a single DataNode
   * C) The client retrieves each block from the appropriate DataNode based on the block locations provided by the NameNode
   * D) The client retrieves the file from the Web Interface
   * **Answer:** C) The client retrieves each block from the appropriate DataNode based on the block locations provided by the NameNode
8. **Which of the following HDFS user commands can be used to list the contents of a directory in HDFS?**
   * A) hdfs dfs -ls
   * B) hdfs dfs -cat
   * C) hdfs fs -list
   * D) hdfs file -show
   * **Answer:** A) hdfs dfs -ls
9. **Which of the following HDFS admin commands is used to check the health of the HDFS filesystem?**
   * A) hdfs fsck /
   * B) hdfs dfsadmin -report
   * C) hdfs fs -status
   * D) hdfs dfs -check
   * **Answer:** B) hdfs dfsadmin -report
10. **What does the hdfs dfs -cat command do in HDFS?**
    * A) It lists the contents of a directory
    * B) It concatenates files and displays them on the console
    * C) It moves files between directories
    * D) It changes the permissions of files
    * **Answer:** B) It concatenates files and displays them on the console
11. **Which HDFS admin command is used to start the NameNode?**
    * A) hdfs start namenode
    * B) start-hadoop namenode
    * C) hadoop namenode start
    * D) hdfs namenode -start
    * **Answer:** C) hadoop namenode start
12. **How can the HDFS Web Interface be used to monitor the cluster’s health?**
    * A) By accessing the URL http://<namenode\_host>:50070/dfshealth.jsp
    * B) By using the hdfs dfsadmin -webhealth command
    * C) By accessing http://<namenode\_host>:8088/cluster
    * D) By opening http://<namenode\_host>:9000/admin
    * **Answer:** A) By accessing the URL http://<namenode\_host>:50070/dfshealth.jsp
13. **In the HDFS Web Interface, which page provides the status of all DataNodes in the cluster?**
    * A) DataNode Health page
    * B) DFS Health page
    * C) Namenode Status page
    * D) System Overview page
    * **Answer:** B) DFS Health page
14. **What does the HDFS Web Interface allow users to do with their files?**
    * A) Directly edit the content of files stored in HDFS
    * B) Upload and download files to and from HDFS
    * C) Execute MapReduce jobs on files
    * D) Change the file system block size
    * **Answer:** B) Upload and download files to and from HDFS
15. **How can a user upload a file from the local system to HDFS using the hdfs dfs command?**
    * A) hdfs dfs -upload <local\_file> <hdfs\_path>
    * B) hdfs dfs -put <local\_file> <hdfs\_path>
    * C) hdfs dfs -copy <local\_file> <hdfs\_path>
    * D) hdfs dfs -push <local\_file> <hdfs\_path>
    * **Answer:** B) hdfs dfs -put <local\_file> <hdfs\_path>
16. **In HDFS, when a file is written, which of the following describes the flow of operations?**
    * A) The file is split into blocks, and the client writes data directly to the DataNodes without the NameNode’s involvement
    * B) The client communicates with the NameNode, which selects DataNodes to store blocks, and the client writes data to these DataNodes
    * C) The file is first encrypted by the client, then written to DataNodes
    * D) The file is stored as a whole on a single DataNode, without replication
    * **Answer:** B) The client communicates with the NameNode, which selects DataNodes to store blocks, and the client writes data to these DataNodes
17. **Which of the following HDFS admin commands is used to change the replication factor of a file?**
    * A) hdfs dfs -setrep
    * B) hdfs dfs -change-replication
    * C) hdfs dfs -update-replication
    * D) hdfs dfs -setblocksize
    * **Answer:** A) hdfs dfs -setrep
18. **What is the default behavior of HDFS replication when a DataNode fails?**
    * A) The data on the failed DataNode is lost permanently
    * B) The data is replicated to another available DataNode to meet the replication factor
    * C) The NameNode stops assigning blocks to the DataNode
    * D) HDFS will mark the block as “under-replicated” but will not replicate until the DataNode recovers
    * **Answer:** B) The data is replicated to another available DataNode to meet the replication factor
19. **Which of the following is responsible for maintaining the namespace of HDFS, including the directory structure and file names?**

* A) DataNode
* B) NameNode
* C) ResourceManager
* D) JobTracker
* **Answer:** B) NameNode

1. **In HDFS, how is data replicated across DataNodes to ensure fault tolerance?**

* A) Data is mirrored between DataNodes every minute
* B) Each block is replicated across multiple DataNodes
* C) Data is written to one DataNode and then replicated manually
* D) Replication is not supported in HDFS
* **Answer:** B) Each block is replicated across multiple DataNodes

1. **When a client requests to read a file in HDFS, what happens first?**

* A) The client reads the file directly from the NameNode
* B) The client asks the NameNode for the block locations
* C) The client retrieves data from a DataNode without involving the NameNode
* D) The client requests the file content from the Web Interface
* **Answer:** B) The client asks the NameNode for the block locations

1. **Which of the following is NOT a valid HDFS command to interact with files?**

* A) hdfs dfs -ls
* B) hdfs dfs -mkdir
* C) hdfs dfs -rm
* D) hdfs dfs -append
* **Answer:** D) hdfs dfs -append

1. **Which command would be used to check the health of the HDFS filesystem?**

* A) hdfs dfsadmin -checkstatus
* B) hdfs fsck /
* C) hdfs dfsadmin -health
* D) hdfs dfs -report
* **Answer:** B) hdfs fsck /

1. **What is the function of the Secondary NameNode in HDFS?**

* A) It stores all the data and acts as a backup for DataNodes
* B) It periodically merges the NameNode's edits with the fsimage file
* C) It handles client requests when the NameNode is unavailable
* D) It manages block replication across DataNodes
* **Answer:** B) It periodically merges the NameNode's edits with the fsimage file

1. **In HDFS, which of the following would you use to remove a file from HDFS?**

* A) hdfs dfs -del
* B) hdfs dfs -rm
* C) hdfs dfs -delete
* D) hdfs dfs -remove
* **Answer:** B) hdfs dfs -rm

1. **Which of the following HDFS user commands is used to view the content of a file stored in HDFS?**

* A) hdfs dfs -view
* B) hdfs dfs -cat
* C) hdfs dfs -show
* D) hdfs dfs -display
* **Answer:** B) hdfs dfs -cat

1. **What does HDFS fsck (File System Check) command do?**

* A) It checks the network connectivity between DataNodes
* B) It verifies the integrity of HDFS and reports on block health
* C) It repairs corrupted files in HDFS
* D) It optimizes HDFS for better performance
* **Answer:** B) It verifies the integrity of HDFS and reports on block health

1. **Which HDFS admin command is used to view the health of the DataNodes in the cluster?**

* A) hdfs dfsadmin -status
* B) hdfs dfsadmin -nodes
* C) hdfs dfsadmin -report
* D) hdfs dfs -health
* **Answer:** C) hdfs dfsadmin -report

1. **In the HDFS Web Interface, which section provides detailed information about active and inactive DataNodes?**

* A) DataNode Health Page
* B) DFS Health Page
* C) Namenode Overview Page
* D) DataNode Status Page
* **Answer:** A) DataNode Health Page

1. **Which of the following is the default port number for accessing the HDFS Web Interface?**

* A) 8080
* B) 50070
* C) 9000
* D) 50075
* **Answer:** B) 50070

1. **Which HDFS user command is used to copy a file from HDFS to the local filesystem?**

* A) hdfs dfs -get
* B) hdfs dfs -put
* C) hdfs dfs -copyToLocal
* D) hdfs dfs -download
* **Answer:** A) hdfs dfs -get

1. **In HDFS, what happens when a DataNode fails?**

* A) The data is lost, and the file becomes unavailable
* B) The NameNode replicates the lost blocks to other available DataNodes
* C) The client must restart the file writing process
* D) The DataNode automatically replicates the blocks to another DataNode
* **Answer:** B) The NameNode replicates the lost blocks to other available DataNodes

1. **Which HDFS admin command is used to change the replication factor of a file in HDFS?**

* A) hdfs dfs -change-replication
* B) hdfs dfs -setrep
* C) hdfs dfs -update-replication
* D) hdfs dfs -set
* **Answer:** B) hdfs dfs -setrep

1. **Which of the following is NOT part of HDFS architecture?**

* A) DataNodes
* B) NameNode
* C) ResourceManager
* D) Secondary NameNode
* **Answer:** C) ResourceManager

1. **Which HDFS component ensures data availability in case of Node failure?**

* A) DataNode replication
* B) NameNode backup
* C) Secondary NameNode checkpointing
* D) Rack-awareness
* **Answer:** A) DataNode replication

1. **What is the primary function of the NameNode in HDFS?**

* A) To store and retrieve file data
* B) To handle read/write requests from clients
* C) To maintain metadata and directory structure
* D) To ensure that all files are compressed
* **Answer:** C) To maintain metadata and directory structure

1. **What happens when a file block in HDFS is under-replicated?**

* A) The file is deleted automatically
* B) HDFS initiates block replication to meet the required replication factor
* C) The block is marked as corrupted
* D) The file is locked and cannot be accessed
* **Answer:** B) HDFS initiates block replication to meet the required replication factor

1. **What is the role of the DataNode in HDFS?**

* A) To manage the block replication and storage
* B) To handle metadata operations
* C) To provide network connectivity for file transfer
* D) To store file directory structure
* **Answer:** A) To manage the block replication and storage

1. **What command is used to view the file block locations in HDFS?**

* A) hdfs dfs -blockinfo
* B) hdfs dfs -stat
* C) hdfs dfs -du
* D) hdfs fsck /
* **Answer:** A) hdfs dfs -blockinfo

1. **Which of the following is the default HDFS command for viewing the file system information?**

* A) hdfs fs -df
* B) hdfs dfs -ls
* C) hdfs dfs -du
* D) hdfs dfsadmin -report
* **Answer:** A) hdfs fs -df

1. **What action does the hdfs dfsadmin -fsck command perform?**

* A) It forces a file check to verify the integrity of the HDFS filesystem
* B) It creates a new file in HDFS
* C) It provides the list of files on HDFS
* D) It checks the memory status of DataNodes
* **Answer:** A) It forces a file check to verify the integrity of the HDFS filesystem

1. **Which component of HDFS is responsible for file replication management and rebalancing?**

* A) NameNode
* B) DataNode
* C) ResourceManager
* D) Secondary NameNode
* **Answer:** A) NameNode

1. **How does HDFS handle data locality?**

* A) It replicates data across all DataNodes
* B) It tries to store data close to the computation resource for better performance
* C) It stores data in remote DataNodes to ensure redundancy
* D) Data is evenly distributed across all available DataNodes
* **Answer:** B) It tries to store data close to the computation resource for better performance

1. **What is the purpose of rack awareness in HDFS?**

* A) To optimize data transfer between racks
* B) To ensure that replicas are distributed across different racks to reduce the risk of data loss
* C) To monitor the health of each rack
* D) To ensure data availability in the event of a rack failure
* **Answer:** B) To ensure that replicas are distributed across different racks to reduce the risk of data loss

1. **Which of the following is the correct command for copying a file from a local system to HDFS?**

* A) hdfs dfs -copyLocal
* B) hdfs dfs -put
* C) hdfs dfs -upload
* D) hdfs dfs -store
* **Answer:** B) hdfs dfs -put

1. **Which of the following is a function of HDFS’s Web Interface?**

* A) Display the block and replica status of files
* B) Display MapReduce job logs
* C) Display the content of files
* D) Provide APIs for DataNode interaction
* **Answer:** A) Display the block and replica status of files

##########################################################################

Session 9

**Getting in touch with Map Reduce Framework**

o Hadoop Map Reduce paradigm,

o Stages of MapReduce

o Map and Reduce tasks,

o Map Reduce Execution Framework,

o Anatomy of a Map Reduce Job run

### **1. What is the primary purpose of the MapReduce framework in Hadoop?**

* A) To store large volumes of data efficiently
* B) To process large datasets in parallel across multiple nodes
* C) To manage cluster resources
* D) To handle data replication and fault tolerance
* **Answer:** B) To process large datasets in parallel across multiple nodes

### **2. In MapReduce, what happens during the map phase?**

* A) Data is divided into smaller blocks
* B) The input data is transformed into a key-value pair
* C) The output data is sorted and reduced
* D) Data is written to the HDFS storage
* **Answer:** B) The input data is transformed into a key-value pair

### **3. Which of the following defines the role of the reduce phase in MapReduce?**

* A) It processes input data and transforms it into key-value pairs
* B) It groups all intermediate data based on the key and performs aggregation
* C) It stores the output data back into HDFS
* D) It splits data into blocks for parallel processing
* **Answer:** B) It groups all intermediate data based on the key and performs aggregation

### **4. What are the main stages of a MapReduce job?**

* A) Input reading, Map phase, Shuffle phase, Reduce phase, Output writing
* B) Data generation, Sorting, Filtering, Reduce phase
* C) Data compression, Mapping, Reducing, Writing output
* D) Input reading, Map phase, Filter phase, Write to file system
* **Answer:** A) Input reading, Map phase, Shuffle phase, Reduce phase, Output writing

### **5. Which component is responsible for splitting input data into chunks during MapReduce execution?**

* A) Mapper
* B) JobTracker
* C) HDFS
* D) InputFormat
* **Answer:** D) InputFormat

### **6. In the MapReduce paradigm, the Mapper output is in which format?**

* A) Text data
* B) Key-Value pair
* C) Compressed binary data
* D) Sorted output
* **Answer:** B) Key-Value pair

### **7. The shuffle phase in MapReduce refers to the process of:**

* A) Merging sorted data from the map output
* B) Sorting the final output before writing it to HDFS
* C) Splitting the input data into smaller blocks
* D) Running the reduce tasks
* **Answer:** A) Merging sorted data from the map output

### **8. Which Hadoop component is responsible for tracking the execution of MapReduce tasks and monitoring their status?**

* A) NameNode
* B) JobTracker
* C) ResourceManager
* D) DataNode
* **Answer:** B) JobTracker

### **9. The Reduce function in MapReduce performs which of the following?**

* A) Sorting input data
* B) Aggregating or reducing the data based on keys
* C) Splitting input data into smaller chunks
* D) Writing the final output to HDFS
* **Answer:** B) Aggregating or reducing the data based on keys

### **10. In MapReduce, the JobConf class is used for:**

* A) Storing configuration properties for the job
* B) Splitting the input data into chunks
* C) Managing job tracking and execution
* D) Defining the output data format
* **Answer:** A) Storing configuration properties for the job

### **11. In the context of MapReduce execution, the JobTracker is responsible for:**

* A) Managing resource allocation and scheduling jobs
* B) Maintaining the directory structure in HDFS
* C) Splitting the input data into blocks
* D) Writing output data to HDFS
* **Answer:** A) Managing resource allocation and scheduling jobs

### **12. How are intermediate key-value pairs between the Map and Reduce stages handled?**

* A) They are immediately written to HDFS
* B) They are passed through a **shuffle and sort** phase
* C) They are discarded after the map phase
* D) They are aggregated in memory before being passed to reducers
* **Answer:** B) They are passed through a **shuffle and sort** phase

### **13. What happens if a MapReduce job fails?**

* A) The entire job is re-run from scratch
* B) The failed task is rescheduled on a different node
* C) The output data is deleted and job execution is stopped
* D) The reduce phase continues while the map phase is restarted
* **Answer:** B) The failed task is rescheduled on a different node

### **14. Which of the following is used to define how MapReduce jobs are configured in Hadoop?**

* A) JobTracker
* B) MapReduce API
* C) JobConf
* D) InputFormat
* **Answer:** C) JobConf

### **15. In the MapReduce framework, which class is used to define the Mapper class?**

* A) Mapper
* B) MapReduceTask
* C) MapperTask
* D) MapProcessor
* **Answer:** A) Mapper

### **16. In a MapReduce job, which component writes the final output to HDFS?**

* A) Mapper
* B) Reducer
* C) JobTracker
* D) OutputFormat
* **Answer:** D) OutputFormat

### **\*\*17. In a MapReduce job, what is the primary purpose of the Combiner?**

* A) To aggregate results at the mapper side before passing them to the reducer
* B) To aggregate the final results after all the reducers have finished
* C) To shuffle the data between mappers and reducers
* D) To validate the data being passed from the mapper to the reducer
* **Answer:** A) To aggregate results at the mapper side before passing them to the reducer

### **18. What is the purpose of the IdentityMapper in MapReduce?**

* A) To perform no transformations and pass input data directly to the reducer
* B) To filter the input data before passing it to the reducer
* C) To aggregate the input data before passing it to the reducer
* D) To change the format of the input data
* **Answer:** A) To perform no transformations and pass input data directly to the reducer

### **19. In the MapReduce execution framework, which component is responsible for resource management and task scheduling?**

* A) JobTracker
* B) TaskTracker
* C) ResourceManager
* D) DataNode
* **Answer:** C) ResourceManager

### **20. What is the role of the TaskTracker in the MapReduce framework?**

* A) To manage job submission and resource allocation
* B) To manage the execution of Map and Reduce tasks on nodes
* C) To monitor the overall health of the cluster
* D) To handle input data splitting and block management
* **Answer:** B) To manage the execution of Map and Reduce tasks on nodes

### **21. The InputFormat in MapReduce defines:**

* A) The output format for writing data
* B) How input data is read and split for processing
* C) How data is shuffled between the map and reduce phases
* D) The compression technique used for input data
* **Answer:** B) How input data is read and split for processing

### **22. What happens in the shuffle phase of MapReduce?**

* A) The data is divided into blocks for parallel processing
* B) The output from all mappers is sorted and grouped by key
* C) Data is aggregated by the mapper before it is passed to reducers
* D) The final output is written to HDFS
* **Answer:** B) The output from all mappers is sorted and grouped by key

### **23. Which of the following classes is used to define the Reducer class in MapReduce?**

* A) Reducer
* B) ReduceTask
* C) MapReduceJob
* D) JobReduce
* **Answer:** A) Reducer

### **24. In the context of MapReduce, which phase involves writing output data to HDFS?**

* A) Map phase
* B) Shuffle phase
* C) Reduce phase
* D) Output phase
* **Answer:** D) Output phase

### **25. Which of the following best describes the anatomy of a MapReduce job?**

* A) Input data is read, mapped, shuffled, reduced, and output data is written
* B) Input data is split, mapped, shuffled, filtered, and then written
* C) Data is first aggregated, then mapped and reduced, and finally written
* D) Input data is sorted, mapped, reduced, and written to a distributed file system
* **Answer:** A) Input data is read, mapped, shuffled, reduced, and output data is written

### **26. How does the MapReduce framework handle failure recovery?**

* A) It automatically retries failed tasks on another node
* B) It stops the job and requires manual intervention
* C) It only retries tasks that fail during the reduce phase
* D) It deletes the entire job and starts over
* **Answer:** A) It automatically retries failed tasks on another node

### **27. What is the default number of reducers in a MapReduce job if not specified?**

* A) 1
* B) 2
* C) 3
* D) It is automatically determined by the system
* **Answer:** A) 1

### **28. Which of the following describes the output of the MapReduce job?**

* A) Key-value pairs generated by the map function
* B) Aggregated key-value pairs produced by the reduce function
* C) Text files containing key-value pairs written to HDFS
* D) The final aggregated output, written as a single file
* **Answer:** C) Text files containing key-value pairs written to HDFS

### **29. What does the map phase of MapReduce do to the input data?**

* A) It sorts the input data
* B) It groups the data based on keys
* C) It filters the data
* D) It splits and processes the input data into key-value pairs
* **Answer:** D) It splits and processes the input data into key-value pairs

### **30. Which of the following is true about MapReduce scalability?**

* A) MapReduce jobs can scale to process petabytes of data across many nodes
* B) MapReduce is not scalable and only works with small data sizes
* C) MapReduce can process large data but only works on a single node
* D) MapReduce only works with structured data
* **Answer:** A) MapReduce jobs can scale to process petabytes of data across many nodes

### **31. What does the map function in MapReduce output?**

* A) A sorted list of the entire input dataset
* B) Key-value pairs that will be processed by the reduce function
* C) A single output file containing the entire dataset
* D) The final aggregated output after performing computations
* **Answer:** B) Key-value pairs that will be processed by the reduce function

### **32. Which phase occurs immediately after the map phase in MapReduce?**

* A) Reduce phase
* B) Shuffle phase
* C) Output phase
* D) Sorting phase
* **Answer:** B) Shuffle phase

### **33. How is the output of a reducer written to HDFS?**

* A) Directly by the reduce function using the output.collect() method
* B) Through the Hadoop Job API’s OutputFormat class
* C) By copying the output files from local storage to HDFS
* D) Using the JobTracker to store the results
* **Answer:** B) Through the Hadoop Job API’s OutputFormat class

### **34. Which of the following is true about the parallelism in MapReduce jobs?**

* A) MapReduce jobs always run sequentially and are never parallelized
* B) The **map phase** is inherently parallel, while the reduce phase is not
* C) Both the map and reduce phases are parallelizable and distributed
* D) MapReduce jobs cannot be parallelized
* **Answer:** C) Both the map and reduce phases are parallelizable and distributed

### **35. In MapReduce, which is a key feature of the shuffle phase?**

* A) Merging all intermediate outputs from the mappers
* B) Aggregating data at the mapper before it is sent to the reducer
* C) Sorting and grouping key-value pairs to prepare them for the reduce phase
* D) Writing output data to the file system
* **Answer:** C) Sorting and grouping key-value pairs to prepare them for the reduce phase

### **36. What is the default behavior when a MapReduce job finishes?**

* A) The job continues processing until it reaches a specific output size
* B) The job outputs the results directly into a database
* C) The job outputs the results to a specified output path in HDFS
* D) The job halts, and no data is written
* **Answer:** C) The job outputs the results to a specified output path in HDFS

### **37. Which of the following is a built-in InputFormat in MapReduce?**

* A) TextInputFormat
* B) SequenceFileInputFormat
* C) KeyValueTextInputFormat
* D) All of the above
* **Answer:** D) All of the above

### **38. Which of the following is NOT a feature of Hadoop MapReduce?**

* A) Fault tolerance through task retries
* B) Support for sorting and filtering data
* C) Support for handling unstructured data only
* D) Automatic parallel processing of tasks
* **Answer:** C) Support for handling unstructured data only

### **39. In a MapReduce job, what is the purpose of InputSplit?**

* A) It divides the input data into manageable chunks for parallel processing
* B) It aggregates data before it is processed by the reducer
* C) It combines the output of multiple mappers into one key-value pair
* D) It defines the format of the output data
* **Answer:** A) It divides the input data into manageable chunks for parallel processing

### **40. What is the role of TaskTracker in MapReduce?**

* A) To track the progress of jobs and report to the JobTracker
* B) To split input data into blocks and manage job execution
* C) To define the output format for the MapReduce job
* D) To provide the user interface for MapReduce job submission
* **Answer:** A) To track the progress of jobs and report to the JobTracker

### **41. Which of the following is a MapReduce data type for representing key-value pairs?**

* A) PairList
* B) Tuple
* C) KeyValue
* D) KeyValuePair
* **Answer:** D) KeyValuePair

### **42. How does MapReduce ensure fault tolerance during job execution?**

* A) By copying data across multiple DataNodes
* B) By retrying failed tasks and re-executing the failed part of the job
* C) By dividing the input into smaller chunks for better distribution
* D) By storing intermediate results in memory
* **Answer:** B) By retrying failed tasks and re-executing the failed part of the job

### **43. Which of the following is true about combiner in MapReduce?**

* A) Combiner is a specialized form of the reduce function applied at the mapper side
* B) Combiner performs the final aggregation after the reduce phase
* C) Combiner aggregates data after both map and reduce phases are completed
* D) Combiner is optional and helps reduce the size of the intermediate data transferred to reducers
* **Answer:** D) Combiner is optional and helps reduce the size of the intermediate data transferred to reducers

### **44. In a MapReduce job, what happens if the input data size is smaller than the number of mappers?**

* A) The job execution will fail
* B) The number of mappers will be reduced to match the input size
* C) The mappers will run in parallel for all data but process less data
* D) The reduce tasks will execute without the map tasks
* **Answer:** C) The mappers will run in parallel for all data but process less data

### **45. What happens when a MapReduce job encounters a duplicate key?**

* A) The duplicate key is discarded
* B) The duplicate key is merged automatically by the reduce function
* C) The job fails and stops processing
* D) The duplicate key is sent to a separate reduce task for further processing
* **Answer:** B) The duplicate key is merged automatically by the reduce function

### **46. What determines the number of reducers in a MapReduce job?**

* A) The size of the input data
* B) The number of mappers in the job
* C) The complexity of the map and reduce functions
* D) The configuration specified by the user or system defaults
* **Answer:** D) The configuration specified by the user or system defaults

### **47. Which of the following is true about output data format in MapReduce?**

* A) It defines how the output data is serialized and written to HDFS
* B) It is always in the form of a text file
* C) It is automatically determined based on the job configuration
* D) Output format is defined in the mapper class itself
* **Answer:** A) It defines how the output data is serialized and written to HDFS

### **48. Which of the following is a common use case for MapReduce?**

* A) File compression
* B) Data analysis on large-scale datasets
* C) Image rendering
* D) Real-time streaming data processing
* **Answer:** B) Data analysis on large-scale datasets

### **49. What is the default number of reducers in a MapReduce job if not specified?**

* A) 0
* B) 1
* C) 2
* D) It is automatically calculated based on the input size
* **Answer:** B) 1

### **50. How does MapReduce optimize performance when handling large datasets?**

* A) By processing all the data on a single node
* B) By distributing data across multiple nodes and parallelizing tasks
* C) By using high-memory machines to process large datasets
* D) By writing intermediate results to disk instead of memory
* **Answer:** B) By distributing data across multiple nodes and parallelizing tasks

##############################################################################

**Session 10**

o YARN Architecture

o YARN Resource Management

o Hadoop Schedulers

o Upgrading cluster from Hadoop1 to Hadoop2

o MapReduce job workflow on YARN

o Migration from MRv1 to MRv2 on YARN : Configuration changes in files

### **Easy Questions (10)**

1. **What does YARN stand for in Hadoop?**
   * A) Yet Another Resource Negotiator
   * B) Your Allocated Resource Network
   * C) You Always Require Nodes
   * D) Yoked Aggregation Resource Nodes
   * **Answer:** A) Yet Another Resource Negotiator
2. **Which component manages resources in YARN?**
   * A) ResourceManager
   * B) NodeManager
   * C) ApplicationMaster
   * D) JobTracker
   * **Answer:** A) ResourceManager
3. **What is the purpose of the ResourceManager in YARN?**
   * A) It executes MapReduce jobs
   * B) It allocates resources for applications in the cluster
   * C) It monitors the health of the Hadoop nodes
   * D) It manages the job history
   * **Answer:** B) It allocates resources for applications in the cluster
4. **Which is the main role of the NodeManager in YARN?**
   * A) Managing the job submission process
   * B) Managing resources across the entire cluster
   * C) Managing the resources on individual nodes and launching containers
   * D) Aggregating job logs
   * **Answer:** C) Managing the resources on individual nodes and launching containers
5. **In YARN, what is the ApplicationMaster responsible for?**
   * A) Managing the physical hardware resources of the cluster
   * B) Running and monitoring a specific application job
   * C) Launching and managing containers for job execution
   * D) Collecting logs from all running applications
   * **Answer:** B) Running and monitoring a specific application job
6. **What feature of YARN provides multitenancy?**
   * A) By allowing multiple jobs to run concurrently in containers
   * B) By providing storage for different users' data
   * C) By ensuring that only one job runs on a single node
   * D) By restricting access to the cluster
   * **Answer:** A) By allowing multiple jobs to run concurrently in containers
7. **Which of the following is NOT a supported YARN Scheduler?**
   * A) FIFO Scheduler
   * B) Capacity Scheduler
   * C) Fair Scheduler
   * D) RoundRobin Scheduler
   * **Answer:** D) RoundRobin Scheduler
8. \*\*What is the \*\*default **scheduler used in YARN?**
   * A) FIFO Scheduler
   * B) Capacity Scheduler
   * C) Fair Scheduler
   * D) Hybrid Scheduler
   * **Answer:** A) FIFO Scheduler
9. **What is the main goal of the YARN ResourceManager?**
   * A) To run MapReduce jobs on the cluster
   * B) To track job status and progress
   * C) To manage resource allocation and scheduling of jobs
   * D) To store input/output data on HDFS
   * **Answer:** C) To manage resource allocation and scheduling of jobs
10. **Which component handles job scheduling in YARN?**

* A) ResourceManager
* B) JobTracker
* C) TaskScheduler
* D) NodeManager
* **Answer:** A) ResourceManager

### **Intermediate Questions (10)**

1. **In Hadoop 2.x, what major component was introduced that improved resource management?**

* A) MapReduce v2 (MRv2)
* B) YARN (Yet Another Resource Negotiator)
* C) Hadoop Distributed File System (HDFS)
* D) Zookeeper
* **Answer:** B) YARN (Yet Another Resource Negotiator)

1. **How does YARN handle job scheduling?**

* A) By allocating jobs to the node with the highest CPU
* B) By using **JobScheduler** for resource allocation
* C) By using the ResourceManager and configured schedulers
* D) Through manual configuration by system administrators
* **Answer:** C) By using the ResourceManager and configured schedulers

1. **Which component of YARN is responsible for managing resources across all nodes in the cluster?**

* A) ApplicationMaster
* B) NodeManager
* C) ResourceManager
* D) JobTracker
* **Answer:** C) ResourceManager

1. **How does YARN handle node failures?**

* A) It automatically retries the job on another node
* B) It ignores the failure and proceeds with the job
* C) It halts all jobs and notifies the administrator
* D) It restarts the entire cluster
* **Answer:** A) It automatically retries the job on another node

1. **What is the role of the Scheduler in YARN?**

* A) To allocate resources to applications running in the cluster
* B) To track job progress
* C) To handle job failures and retries
* D) To monitor cluster health
* **Answer:** A) To allocate resources to applications running in the cluster

1. **In Hadoop 1.x, which component handled both resource management and job scheduling?**

* A) ResourceManager
* B) JobTracker
* C) TaskTracker
* D) NodeManager
* **Answer:** B) JobTracker

1. **Which configuration file in YARN defines the ResourceManager address?**

* A) core-site.xml
* B) hdfs-site.xml
* C) yarn-site.xml
* D) mapred-site.xml
* **Answer:** C) yarn-site.xml

1. **In YARN, how are jobs assigned to nodes?**

* A) Based on task priority and cluster utilization
* B) Based on user configuration and job size
* C) Based on job type and task availability
* D) Based on system resource constraints and the scheduler's policy
* **Answer:** D) Based on system resource constraints and the scheduler's policy

1. **What happens during the container allocation process in YARN?**

* A) YARN reserves the entire node for the job
* B) ResourceManager allocates specific resources like CPU, memory, and disk to a container
* C) NodeManager picks the best available node
* D) Applications are queued until there are available resources
* **Answer:** B) ResourceManager allocates specific resources like CPU, memory, and disk to a container

1. \*\*What is the \*\*default **scheduler in YARN used for?**

* A) Allocating jobs in a First In, First Out (FIFO) manner
* B) Scheduling jobs based on job priority
* C) Ensuring the jobs share resources equally
* D) Allocating resources based on availability in the cluster
* **Answer:** A) Allocating jobs in a First In, First Out (FIFO) manner

### **Hard Questions (10)**

1. **Which of the following is NOT a feature of YARN?**

* A) Dynamic resource allocation
* B) Support for both batch and real-time processing
* C) Ability to run non-MapReduce applications
* D) Centralized data storage for job results
* **Answer:** D) Centralized data storage for job results

1. \*\*What is the **role of the NodeManager in handling failures in YARN?**

* A) It retries the task within the same container
* B) It notifies the ResourceManager of failures
* C) It restarts the entire cluster
* D) It handles job resubmission on available nodes
* **Answer:** D) It handles job resubmission on available nodes

1. **In YARN, which of the following actions requires administrative intervention?**

* A) Adjusting the memory allocation for containers
* B) Changing the ResourceManager’s scheduler
* C) Allocating resources to a MapReduce job
* D) Automatically scaling the cluster
* **Answer:** B) Changing the ResourceManager’s scheduler

1. **In a YARN-managed cluster, which resource is NOT allocated by the ResourceManager?**

* A) CPU cores
* B) Network bandwidth
* C) Memory
* D) Disk space
* **Answer:** B) Network bandwidth

1. \*\*How does **YARN handle resource contention when multiple applications request resources simultaneously?**

* A) It allocates resources based on job priority
* B) It cancels lower-priority jobs
* C) It uses the scheduler to divide resources fairly or as per defined policies
* D) It automatically allocates resources based on CPU usage
* **Answer:** C) It uses the scheduler to divide resources fairly or as per defined policies

1. **What is the major difference between MRv1 and MRv2 in YARN?**

* A) MRv2 provides an interactive UI for users
* B) MRv2 uses YARN for resource management instead of JobTracker
* C) MRv1 supports only batch processing jobs
* D) MRv1 can only run on single-node clusters
* **Answer:** B) MRv2 uses YARN for resource management instead of JobTracker

1. **Which of the following is a YARN Scheduler that prioritizes capacity for large applications?**

* A) FIFO Scheduler
* B) Fair Scheduler
* C) Capacity Scheduler
* D) Dynamic Scheduler
* **Answer:** C) Capacity Scheduler

1. \*\*In YARN, what is the **role of the ApplicationMaster during job execution?**

* A) To manage resource allocation across the entire cluster
* B) To request resources and monitor the execution of tasks within containers
* C) To manage node failures and reschedule jobs
* D) To monitor system performance and report issues
* **Answer:** B) To request resources and monitor the execution of tasks within containers

1. **In the Capacity Scheduler of YARN, what does the "resource partition" mean?**

* A) The division of available resources into pools for different applications or users
* B) The allocation of nodes for specific jobs based on job size
* C) The number of available nodes in the cluster
* D) The minimum resource requirement for each job
* **Answer:** A) The division of available resources into pools for different applications or users

1. \*\*When upgrading from \*\*Hadoop 1.x to 2.x, what is a key change in how **MapReduce jobs are executed?**

* A) MapReduce jobs are no longer supported in Hadoop 2.x
* B) MapReduce jobs are scheduled through YARN
* C) YARN is used only for resource management, not job execution
* D) The JobTracker is replaced by the ResourceManager in YARN
* **Answer:** B) MapReduce jobs are scheduled through YARN

### **Easy Questions (7)**

1. **Which YARN component is responsible for managing individual node resources?**
   * A) ResourceManager
   * B) NodeManager
   * C) JobTracker
   * D) ApplicationMaster
   * **Answer:** B) NodeManager
2. **Which YARN component is responsible for job execution and tracking?**
   * A) ResourceManager
   * B) NodeManager
   * C) ApplicationMaster
   * D) TaskTracker
   * **Answer:** C) ApplicationMaster
3. **Which of the following can run on YARN apart from MapReduce?**
   * A) Spark
   * B) Pig
   * C) Hive
   * D) All of the above
   * **Answer:** D) All of the above
4. **What is the role of Container in YARN?**
   * A) It stores the data produced by MapReduce tasks
   * B) It isolates jobs running in the cluster
   * C) It contains the computational resources required for job execution
   * D) It contains job logs for later use
   * **Answer:** C) It contains the computational resources required for job execution
5. **What is the default scheduling policy in YARN for job execution?**
   * A) Fair Scheduling
   * B) FIFO (First In First Out)
   * C) Dynamic Scheduling
   * D) Capacity Scheduling
   * **Answer:** B) FIFO (First In First Out)
6. **Which configuration file in YARN defines the ResourceManager hostname?**
   * A) core-site.xml
   * B) hdfs-site.xml
   * C) yarn-site.xml
   * D) mapred-site.xml
   * **Answer:** C) yarn-site.xml
7. **Which of the following statements is true for YARN?**
   * A) YARN is used for managing resources in a Hadoop cluster
   * B) YARN only supports MapReduce jobs
   * C) YARN is not scalable
   * D) YARN is only useful for single-node clusters
   * **Answer:** A) YARN is used for managing resources in a Hadoop cluster

### **Intermediate Questions (8)**

1. **How does YARN handle node failures?**
   * A) It retries tasks on other available nodes
   * B) It stops all tasks in the cluster
   * C) It automatically restarts failed jobs
   * D) It marks nodes as offline
   * **Answer:** A) It retries tasks on other available nodes
2. **Which YARN scheduler provides fair share of resources among all users and jobs?**
   * A) FIFO Scheduler
   * B) Capacity Scheduler
   * C) Fair Scheduler
   * D) Dynamic Scheduler
   * **Answer:** C) Fair Scheduler
3. **What happens when YARN detects a task failure?**

* A) It restarts the entire cluster
* B) It retries the failed task on another node
* C) It ignores the failure and continues with the next task
* D) It cancels the job and notifies the user
* **Answer:** B) It retries the failed task on another node

1. **Which of the following resources is NOT considered for container allocation in YARN?**

* A) CPU
* B) Memory
* C) Disk space
* D) Network bandwidth
* **Answer:** D) Network bandwidth

1. **In YARN, what is ApplicationMaster responsible for during job execution?**

* A) Allocating resources to the job
* B) Monitoring job progress and handling failures
* C) Managing node failures and reallocation of resources
* D) Managing job logs and output storage
* **Answer:** B) Monitoring job progress and handling failures

1. **Which YARN scheduler would you choose if you want to ensure fair allocation of resources to multiple users?**

* A) FIFO Scheduler
* B) Capacity Scheduler
* C) Fair Scheduler
* D) JobScheduler
* **Answer:** C) Fair Scheduler

1. **Which of the following does YARN use to manage large-scale resources across multiple clusters?**

* A) ResourceManager
* B) TaskTracker
* C) NodeManager
* D) ApplicationMaster
* **Answer:** A) ResourceManager

### **Hard Questions (5)**

1. **What is the impact of YARN’s node failure handling mechanism on job execution?**

* A) It stops all jobs and restarts the entire cluster
* B) It retries the tasks from the failed node on another available node
* C) It reassigns the entire job to a different node
* D) It disables job execution until node failure is resolved
* **Answer:** B) It retries the tasks from the failed node on another available node

1. **Which of the following is a key feature of YARN’s Capacity Scheduler?**

* A) Ensures jobs are processed in FIFO order
* B) Allows jobs to share resources equally across all users
* C) Allocates a fixed amount of resources to queues for different applications
* D) Prioritizes jobs based on size and type
* **Answer:** C) Allocates a fixed amount of resources to queues for different applications

1. **Which of the following is NOT an advantage of YARN over MapReduce v1?**

* A) Improved scalability and flexibility
* B) Allows multi-tenancy in Hadoop clusters
* C) Single point of failure for job execution
* D) It decouples resource management and job execution
* **Answer:** C) Single point of failure for job execution

1. **In YARN, what happens if there is insufficient memory for a task in a container?**

* A) The job is automatically terminated
* B) The task is moved to a different container
* C) The task is retried on the same container
* D) The task is paused until memory becomes available
* **Answer:** B) The task is moved to a different container

1. **In YARN, what is the role of Queue Manager in the Capacity Scheduler?**

* A) It manages the overall cluster resources
* B) It allocates resources to each application and defines resource usage policies
* C) It monitors the health of the nodes
* D) It schedules jobs based on the user’s priority
* **Answer:** B) It allocates resources to each application and defines resource usage policies

### **Bonus Questions (5)**

1. **Which file in YARN configuration contains the settings for the ResourceManager's hostname and port?**

* A) core-site.xml
* B) hdfs-site.xml
* C) mapred-site.xml
* D) yarn-site.xml
* **Answer:** D) yarn-site.xml

1. **In YARN, which of the following components can be monitored via the ResourceManager UI?**

* A) Node failures
* B) Application status and resource usage
* C) Job logs
* D) TaskTracker health
* **Answer:** B) Application status and resource usage

1. **In the YARN Capacity Scheduler, what is the function of Resource Allocation?**

* A) To prioritize jobs based on their execution order
* B) To allocate resources to each job based on predefined policies
* C) To allocate resources equally among all jobs
* D) To automatically assign jobs to specific nodes
* **Answer:** B) To allocate resources to each job based on predefined policies

1. \*\*How does **YARN handle scalability in a multi-node cluster?**

* A) By assigning one job per node
* B) By dynamically allocating and releasing resources as required by jobs
* C) By requiring manual intervention to add nodes
* D) By restricting the number of jobs that can run concurrently
* **Answer:** B) By dynamically allocating and releasing resources as required by jobs

1. **In YARN, what happens when a MapReduce job is submitted?**

* A) It is immediately executed by the ResourceManager
* B) It is submitted to the JobTracker for execution
* C) The ApplicationMaster is launched to request resources from the ResourceManager
* D) The job is added to a job queue and waits for an available node
* **Answer:** C) The ApplicationMaster is launched to request resources from the ResourceManager

1. **What is the role of Container in YARN?**

* A) It holds the resources for a single application task
* B) It provides storage for input/output data
* C) It is responsible for job scheduling
* D) It monitors cluster health
* **Answer:** A) It holds the resources for a single application task

#########################################################################

Session 11

**Security in Hadoop**

o HDFS Security Model

o LDAP and Hadoop

o LDAP support in Hadoop

### **Easy Questions (7)**

1. **Which of the following is a common method used to secure data in HDFS?**
   * A) Data encryption
   * B) Data compression
   * C) Data replication
   * D) Data indexing
   * **Answer:** A) Data encryption
2. **Which of the following can be used to authenticate users in Hadoop?**
   * A) Kerberos
   * B) LDAP
   * C) SSL
   * D) All of the above
   * **Answer:** D) All of the above
3. **Which of the following is true about HDFS security?**
   * A) HDFS security is enabled by default
   * B) HDFS uses the Hadoop security manager for authentication
   * C) HDFS does not support encryption
   * D) HDFS uses Kerberos for authentication
   * **Answer:** D) HDFS uses Kerberos for authentication
4. **What is the purpose of Kerberos in Hadoop?**
   * A) Data replication
   * B) Authentication of users and services
   * C) Job scheduling
   * D) Resource management
   * **Answer:** B) Authentication of users and services
5. **In Hadoop, LDAP (Lightweight Directory Access Protocol) is commonly used for**?\*\*
   * A) Data storage
   * B) User authentication and authorization
   * C) Job scheduling
   * D) Resource management
   * **Answer:** B) User authentication and authorization
6. **Which of the following is a requirement for enabling Kerberos authentication in Hadoop?**
   * A) A secure network
   * B) A dedicated Kerberos server
   * C) A centralized password policy
   * D) A backup resource manager
   * **Answer:** B) A dedicated Kerberos server
7. \*\*What is the \*\*role of **Hadoop's NameNode in terms of security?**
   * A) It manages user access controls and permissions
   * B) It handles Kerberos authentication
   * C) It stores and manages metadata for HDFS
   * D) It executes MapReduce jobs
   * **Answer:** A) It manages user access controls and permissions

### **Intermediate Questions (8)**

1. **Which of the following services is used in Hadoop to enable LDAP integration for authentication?**
   * A) Kerberos
   * B) Active Directory
   * C) LDAP server
   * D) Zookeeper
   * **Answer:** C) LDAP server
2. **What is the default behavior of Hadoop in terms of user authorization?**
   * A) Users are granted full access by default
   * B) Users need explicit permissions to access data
   * C) All users are denied access by default
   * D) Hadoop does not support user authorization
   * **Answer:** B) Users need explicit permissions to access data
3. **Which of the following statements is true about Kerberos security in Hadoop?**
   * A) Kerberos security is only used for MapReduce jobs
   * B) Kerberos provides encryption for data during transit
   * C) Kerberos is used only for authentication of the ResourceManager
   * D) Kerberos is used for authenticating users and services in Hadoop
   * **Answer:** D) Kerberos is used for authenticating users and services in Hadoop
4. **What is a key difference between Kerberos authentication and LDAP authentication in Hadoop?**
   * A) Kerberos uses certificates for authentication, while LDAP uses passwords
   * B) Kerberos is used for network encryption, while LDAP is used for user authentication
   * C) Kerberos is a protocol for service authentication, while LDAP is a directory service used for storing user information
   * D) There is no difference
   * **Answer:** C) Kerberos is a protocol for service authentication, while LDAP is a directory service used for storing user information
5. **In Hadoop, which configuration file is commonly used to set Kerberos authentication parameters?**
   * A) core-site.xml
   * B) hdfs-site.xml
   * C) yarn-site.xml
   * D) krb5.conf
   * **Answer:** D) krb5.conf
6. **How does HDFS ensure data confidentiality?**
   * A) By encrypting data at rest
   * B) By enabling password authentication for all users
   * C) By using SSL/TLS for data transmission
   * D) By replicating data across nodes
   * **Answer:** A) By encrypting data at rest
7. **Which of the following authentication methods can be integrated with Hadoop for user access management?**
   * A) LDAP
   * B) Kerberos
   * C) Both LDAP and Kerberos
   * D) OAuth
   * **Answer:** C) Both LDAP and Kerberos
8. **Which Hadoop component can be configured to use SSL for secure communication between client and cluster?**
   * A) ResourceManager
   * B) JobTracker
   * C) NameNode
   * D) Both A and C
   * **Answer:** D) Both A and C
9. **Which of the following does Kerberos security in Hadoop protect against?**
   * A) Data loss
   * B) Unauthorized access
   * C) Data corruption
   * D) Task failures
   * **Answer:** B) Unauthorized access

### **Hard Questions (5)**

1. \*\*What is the role of **Hadoop's Key Management Server (KMS) in security?**
   * A) To manage Kerberos tickets for users
   * B) To provide authorization policies for all Hadoop services
   * C) To manage encryption keys used for encrypting HDFS data
   * D) To monitor access to HDFS
   * **Answer:** C) To manage encryption keys used for encrypting HDFS data
2. **In Hadoop, when LDAP authentication is configured, what must happen for a user to access Hadoop services?**
   * A) The user must be authenticated through Kerberos first
   * B) The user must be part of a specific LDAP group that has been granted access
   * C) The user must authenticate using SSL certificates
   * D) The user must be manually authorized by the administrator
   * **Answer:** B) The user must be part of a specific LDAP group that has been granted access
3. **Which of the following are required steps when enabling Kerberos authentication in Hadoop?**
   * A) Set up a Kerberos Key Distribution Center (KDC)
   * B) Enable Kerberos in the core-site.xml and hdfs-site.xml
   * C) Create a service principal for each Hadoop service
   * D) All of the above
   * **Answer:** D) All of the above
4. **Which of the following security mechanisms can Hadoop integrate to enable role-based access control (RBAC)?**
   * A) Kerberos
   * B) Apache Ranger
   * C) LDAP
   * D) SSL
   * **Answer:** B) Apache Ranger

### **Easy Questions (10)**

1. **Which of the following security protocols is used by Hadoop for authentication?**
   * A) LDAP
   * B) Kerberos
   * C) SSL
   * D) SSH
   * **Answer:** B) Kerberos
2. **What is the default method for user authentication in Hadoop?**
   * A) Simple authentication
   * B) Kerberos authentication
   * C) Token-based authentication
   * D) Password authentication
   * **Answer:** B) Kerberos authentication
3. **Which security protocol is required for HDFS encryption?**
   * A) SSL
   * B) Kerberos
   * C) AES
   * D) SSL/TLS
   * **Answer:** B) Kerberos
4. **Which component of Hadoop handles user authentication using Kerberos?**
   * A) ResourceManager
   * B) NameNode
   * C) Kerberos Key Distribution Center (KDC)
   * D) JobTracker
   * **Answer:** C) Kerberos Key Distribution Center (KDC)
5. **Which of the following is the main purpose of Kerberos authentication in Hadoop?**
   * A) Encrypting data at rest
   * B) Securing network communications
   * C) Authenticating users and services
   * D) Compressing data
   * **Answer:** C) Authenticating users and services
6. **What is the primary function of LDAP in Hadoop security?**
   * A) Encrypting data
   * B) Storing user credentials
   * C) Managing encryption keys
   * D) Managing job scheduling
   * **Answer:** B) Storing user credentials
7. **What happens when Kerberos authentication fails in Hadoop?**
   * A) The job execution fails and is retried
   * B) The user is denied access to Hadoop services
   * C) The cluster shuts down automatically
   * D) The job runs without authentication
   * **Answer:** B) The user is denied access to Hadoop services
8. **Which of the following is the standard security model for Hadoop?**
   * A) SSH-based authentication
   * B) Kerberos-based authentication
   * C) SSL/TLS encryption
   * D) Token-based authentication
   * **Answer:** B) Kerberos-based authentication
9. **What is the default user authentication method used by HDFS?**
   * A) Kerberos authentication
   * B) Simple authentication
   * C) Token-based authentication
   * D) OpenLDAP
   * **Answer:** A) Kerberos authentication
10. **Which of the following can secure Hadoop communications between client and cluster?**
    * A) Kerberos
    * B) SSL/TLS
    * C) VPNs
    * D) Both A and B
    * **Answer:** D) Both A and B

### **Intermediate Questions (10)**

1. **Which of the following is a common feature of LDAP in Hadoop?**
   * A) It is used to encrypt user data
   * B) It stores user credentials in a centralized directory
   * C) It is used only for MapReduce job authentication
   * D) It does not support authentication
   * **Answer:** B) It stores user credentials in a centralized directory
2. **How does Kerberos authentication affect Hadoop services like HDFS?**
   * A) It secures communication between services using SSL
   * B) It adds a layer of security by verifying the identity of users and services
   * C) It automatically encrypts data stored in HDFS
   * D) It improves job scheduling efficiency
   * **Answer:** B) It adds a layer of security by verifying the identity of users and services
3. **Which of the following can be used for role-based access control (RBAC) in Hadoop security?**
   * A) Kerberos
   * B) Apache Ranger
   * C) Apache NiFi
   * D) YARN ResourceManager
   * **Answer:** B) Apache Ranger
4. **Which Hadoop security feature is used to encrypt data stored in HDFS?**
   * A) Apache Ranger
   * B) Kerberos
   * C) HDFS Transparent Data Encryption (TDE)
   * D) SSL/TLS
   * **Answer:** C) HDFS Transparent Data Encryption (TDE)
5. **How does Kerberos protect data integrity in Hadoop?**
   * A) By encrypting data at rest
   * B) By verifying the authenticity of users and services
   * C) By monitoring user activity logs
   * D) By blocking unauthorized IP addresses
   * **Answer:** B) By verifying the authenticity of users and services
6. **In LDAP integration with Hadoop, what does a user group represent?**
   * A) A set of users who share the same encryption keys
   * B) A collection of users who have specific permissions within Hadoop services
   * C) A list of services within the Hadoop ecosystem
   * D) A group of application masters in Hadoop
   * **Answer:** B) A collection of users who have specific permissions within Hadoop services
7. **Which configuration setting in Hadoop is used to specify the Kerberos principal for each service?**
   * A) hadoop.security.authentication
   * B) kerberos.principal
   * C) kerberos.keytab
   * D) hdfs.security.authentication
   * **Answer:** A) hadoop.security.authentication
8. **Which Hadoop component manages security policies across multiple services?**
   * A) ResourceManager
   * B) Apache Ranger
   * C) JobTracker
   * D) DataNode
   * **Answer:** B) Apache Ranger
9. **What does Apache Ranger provide for Hadoop security?**
   * A) Authentication using Kerberos
   * B) Data encryption in HDFS
   * C) Centralized security management and auditing
   * D) Job scheduling capabilities
   * **Answer:** C) Centralized security management and auditing

### **Hard Questions (10)**

1. **When configuring Kerberos authentication in Hadoop, what is a Kerberos principal?**
   * A) A set of credentials used by Hadoop services
   * B) A secure key for encrypting data
   * C) A unique identifier for a user or service in the Kerberos database
   * D) A server that stores Hadoop logs
   * **Answer:** C) A unique identifier for a user or service in the Kerberos database
2. **Which of the following services is not required when setting up Kerberos authentication in Hadoop?**
   * A) Key Distribution Center (KDC)
   * B) Kerberos Ticket Granting Server (TGS)
   * C) ResourceManager
   * D) LDAP server
   * **Answer:** D) LDAP server
3. **Which of the following actions is required to set up Kerberos authentication for Hadoop services?**
   * A) Assign user roles within YARN
   * B) Create service principals for each Hadoop service
   * C) Enable SSL encryption for HDFS
   * D) Use Apache Ranger for auditing
   * **Answer:** B) Create service principals for each Hadoop service
4. **What does HDFS encryption do to the data stored on disk?**
   * A) It prevents unauthorized users from reading the data
   * B) It compresses the data to reduce storage costs
   * C) It replicates data to a backup location
   * D) It ensures high availability of data
   * **Answer:** A) It prevents unauthorized users from reading the data
5. **How does LDAP improve security in Hadoop?**
   * A) By automating job scheduling
   * B) By managing and centralizing user authentication
   * C) By encrypting user data
   * D) By ensuring data consistency across nodes
   * **Answer:** B) By managing and centralizing user authentication
6. **Which of the following security measures can be used to audit access to HDFS files?**
   * A) Apache Ranger
   * B) Kerberos
   * C) SSL/TLS
   * D) YARN ResourceManager
   * **Answer:** A) Apache Ranger
7. **How can Hadoop achieve centralized access control for its services?**
   * A) Using Kerberos for authentication
   * B) Using Apache Ranger for policy management
   * C) Using a custom access control list (ACL)
   * D) Using YARN for resource management
   * **Answer:** B) Using Apache Ranger for policy management
8. **What type of security does Kerberos provide for Hadoop services?**
   * A) Data encryption
   * B) User authentication and service authentication
   * C) Job scheduling
   * D) Data replication
   * **Answer:** B) User authentication and service authentication
9. **Which of the following Hadoop services is not directly impacted by Kerberos authentication?**
   * A) NameNode
   * B) JobTracker
   * C) ResourceManager
   * D) DataNode
   * **Answer:** D) DataNode
10. **What is the role of the Keytab file in Kerberos authentication for Hadoop?**
    * A) It stores user credentials in a secure format
    * B) It allows services to authenticate to the Kerberos server without requiring passwords
    * C) It contains service-level policies for Hadoop jobs
    * D) It is used to monitor Kerberos authentication requests
    * **Answer:** B) It allows services to authenticate to the Kerberos server without requiring passwords

Session 12 :

**Hadoop Cluster Planning**

o Choosing hardware and operating systems,

o OS comparison based on features like kernel tuning, disk swapping & etc.

o Based on scenario and workload identify hardware, cluster size

o Based on scenario identify eco-system components

o Identify key network components, Network topology/design based on network usage in Hadoop

### **Easy Questions (15)**

1. **When choosing hardware for a Hadoop cluster, which of the following is the most important factor?**
   * A) CPU clock speed
   * B) Disk storage and memory
   * C) Display resolution
   * D) Network adapter type
   * **Answer:** B) Disk storage and memory
2. **Which of the following operating systems is commonly used in Hadoop clusters?**
   * A) Windows
   * B) Linux
   * C) macOS
   * D) Android
   * **Answer:** B) Linux
3. **Which of the following is the most critical for Hadoop performance in terms of hardware?**
   * A) Processor speed
   * B) Disk I/O and network bandwidth
   * C) Screen resolution
   * D) Video processing unit (GPU)
   * **Answer:** B) Disk I/O and network bandwidth
4. **Which of the following is not a typical hardware requirement for a Hadoop cluster?**
   * A) Sufficient RAM
   * B) High-performance CPUs
   * C) Multiple network interfaces
   * D) High-end graphics card
   * **Answer:** D) High-end graphics card
5. **When planning the operating system for Hadoop, which feature is most important for system performance?**
   * A) Kernel tuning for better disk I/O
   * B) User interface customization
   * C) Default system themes
   * D) CPU architecture
   * **Answer:** A) Kernel tuning for better disk I/O
6. **Which of the following is important to consider when selecting OS for Hadoop?**
   * A) Compatibility with hardware
   * B) Pre-installed graphical user interface
   * C) Maximum file size support
   * D) Gaming capabilities
   * **Answer:** A) Compatibility with hardware
7. **In Hadoop, which network component is used for managing cluster communication?**
   * A) Network switches
   * B) Servers
   * C) Workstations
   * D) Firewalls
   * **Answer:** A) Network switches
8. **Which type of storage should be selected for Hadoop clusters to achieve optimal performance?**
   * A) Solid-state drives (SSD)
   * B) Optical drives
   * C) Magnetic tape
   * D) Blu-ray discs
   * **Answer:** A) Solid-state drives (SSD)
9. **When planning for Hadoop clusters, which factor influences the choice of hardware configuration the most?**
   * A) Workload size and type
   * B) User interface
   * C) Display settings
   * D) Number of users
   * **Answer:** A) Workload size and type
10. \*\*Which of the following is **true about disk swapping in the context of Hadoop clusters?**
    * A) Disk swapping is used to improve I/O performance
    * B) Disk swapping is generally avoided due to performance degradation
    * C) Disk swapping helps in faster data retrieval
    * D) Hadoop clusters use disk swapping for network optimization
    * **Answer:** B) Disk swapping is generally avoided due to performance degradation
11. **Which of the following components is not a part of Hadoop's ecosystem?**
    * A) HDFS
    * B) YARN
    * C) MapReduce
    * D) Active Directory
    * **Answer:** D) Active Directory
12. **For a Hadoop cluster, which type of network topology is typically recommended for large-scale deployments?**
    * A) Star topology
    * B) Ring topology
    * C) Mesh topology
    * D) Tree topology
    * **Answer:** C) Mesh topology
13. **Which factor is least important when deciding the hardware for a Hadoop cluster?**
    * A) Disk speed and capacity
    * B) Network throughput
    * C) Power supply efficiency
    * D) CPU performance
    * **Answer:** C) Power supply efficiency
14. **When deciding on cluster size for a Hadoop implementation, which factor is most important?**
    * A) Number of physical servers
    * B) Data volume and processing needs
    * C) Number of Hadoop services
    * D) Operating system preference
    * **Answer:** B) Data volume and processing needs
15. **In Hadoop clusters, which component primarily handles job scheduling and resource management?**
    * A) ResourceManager
    * B) DataNode
    * C) NameNode
    * D) JobTracker
    * **Answer:** A) ResourceManager

### **Intermediate Questions (15)**

1. **Which of the following is a best practice when deciding the size of a Hadoop cluster?**
   * A) Always overestimate the hardware to avoid future bottlenecks
   * B) Consider both current and future data growth
   * C) Start with minimum hardware and add as needed
   * D) Only use physical hardware for cluster deployment
   * **Answer:** B) Consider both current and future data growth
2. **Which factor affects network performance in Hadoop clusters the most?**
   * A) CPU speed
   * B) Disk latency
   * C) Network latency and bandwidth
   * D) Number of DataNodes
   * **Answer:** C) Network latency and bandwidth
3. **What is the main consideration when choosing storage for a Hadoop cluster?**
   * A) Cost per terabyte
   * B) Access speed and reliability
   * C) Storage color
   * D) Maximum storage capacity per node
   * **Answer:** B) Access speed and reliability
4. **For a high-performance Hadoop cluster, which network topology should be used for optimal data distribution?**
   * A) Star topology
   * B) Mesh topology
   * C) Ring topology
   * D) Bus topology
   * **Answer:** B) Mesh topology
5. **When selecting a hardware configuration for Hadoop, which component is critical for efficient data processing?**
   * A) RAM and CPU
   * B) GPU
   * C) Network interfaces
   * D) Video cards
   * **Answer:** A) RAM and CPU
6. **When planning a Hadoop cluster, which of the following network components is crucial for ensuring fast communication between nodes?**
   * A) Network cables (high-speed Ethernet)
   * B) Printers
   * C) Router
   * D) Firewalls
   * **Answer:** A) Network cables (high-speed Ethernet)
7. \*\*Which of the following is **required for efficient disk swapping in a Hadoop cluster?**
   * A) Low-latency disk
   * B) Fast CPU cores
   * C) High throughput network
   * D) No disk swapping should be used
   * **Answer:** D) No disk swapping should be used
8. **When choosing the operating system for a Hadoop cluster, which of the following is the most important aspect for performance?**
   * A) Kernel tuning and file system support
   * B) Number of graphical user interfaces
   * C) Pre-installed software packages
   * D) Audio quality
   * **Answer:** A) Kernel tuning and file system support
9. **Which component is responsible for data replication across different nodes in a Hadoop cluster?**
   * A) ResourceManager
   * B) DataNode
   * C) JobTracker
   * D) NameNode
   * **Answer:** D) NameNode
10. **For network design in Hadoop, which type of network topology offers the best performance for large clusters with multiple nodes?**
    * A) Tree topology
    * B) Mesh topology
    * C) Hybrid topology
    * D) Star topology
    * **Answer:** B) Mesh topology
11. **Which Hadoop ecosystem component should be deployed for resource management and task scheduling?**
    * A) HDFS
    * B) YARN
    * C) Zookeeper
    * D) MapReduce
    * **Answer:** B) YARN
12. **When setting up a Hadoop cluster, which of the following factors should be considered when determining cluster size?**
    * A) Cluster hardware size
    * B) Workload type and data processing volume
    * C) Operating system preferences
    * D) Number of available IT personnel
    * **Answer:** B) Workload type and data processing volume
13. **When determining hardware specifications for a Hadoop cluster, what is the primary consideration for processing data efficiently?**
    * A) Disk I/O
    * B) RAM and network speed
    * C) CPU clock speed
    * D) SSDs
    * **Answer:** B) RAM and network speed
14. **Which network component ensures redundancy in a Hadoop cluster by managing data replication?**
    * A) NameNode
    * B) ResourceManager
    * C) DataNode
    * D) Network switch
    * **Answer:** A) NameNode
15. **Which file system is commonly used in Hadoop to store data across multiple nodes in the cluster?**
    * A) NTFS
    * B) HDFS
    * C) ext4
    * D) FAT32
    * **Answer:** B) HDFS

### **Hard Questions (20)**

1. **Which of the following advanced network topologies offers optimal data locality and fault tolerance in a large Hadoop cluster?**
   * A) Hybrid topology
   * B) Mesh topology
   * C) Star topology
   * D) Ring topology
   * **Answer:** A) Hybrid topology
2. **For a Hadoop cluster, when scaling compute capacity, which factor is most critical to balance?**
   * A) Network bandwidth
   * B) Disk storage
   * C) Compute (CPU) resources
   * D) Number of job trackers
   * **Answer:** C) Compute (CPU) resources
3. \*\*Which of the following **best describes the role of Zookeeper in a Hadoop cluster?**
   * A) Data replication
   * B) Coordination and synchronization of cluster services
   * C) Data storage and retrieval
   * D) Job scheduling
   * **Answer:** B) Coordination and synchronization of cluster services
4. **What is the impact of inadequate disk storage when planning a Hadoop cluster?**
   * A) Improved performance
   * B) Increased data replication errors
   * C) Reduced cluster capacity and higher latency
   * D) Faster data processing
   * **Answer:** C) Reduced cluster capacity and higher latency
5. **How does kernel tuning improve the performance of Hadoop in a cluster?**
   * A) It reduces disk swapping and improves network throughput
   * B) It optimizes Hadoop job scheduling
   * C) It improves the ability to run larger datasets
   * D) It makes the system easier to manage
   * **Answer:** A) It reduces disk swapping and improves network throughput
6. **Which component of Hadoop architecture is responsible for determining how much data is replicated across nodes in the cluster?**
   * A) DataNode
   * B) NameNode
   * C) ResourceManager
   * D) JobTracker
   * **Answer:** B) NameNode
7. \*\*In Hadoop, what **is the primary benefit of a highly redundant network topology?**
   * A) Faster job execution times
   * B) Increased cluster availability and fault tolerance
   * C) Improved disk I/O speed
   * D) Lower power consumption
   * **Answer:** B) Increased cluster availability and fault tolerance
8. **In Hadoop, which hardware component is essential for high availability of data in the cluster?**
   * A) Data replication across multiple nodes
   * B) High-speed network adapters
   * C) Solid-state drives
   * D) High-performance CPUs
   * **Answer:** A) Data replication across multiple nodes
9. **When planning the storage architecture for a Hadoop cluster, which type of file system is typically considered for high throughput?**
   * A) HDFS
   * B) ext4
   * C) ZFS
   * D) Btrfs
   * **Answer:** A) HDFS
10. **In terms of security, what is the most critical aspect when planning a Hadoop cluster?**
    * A) Kerberos authentication
    * B) CPU security
    * C) File permissions
    * D) Data replication
    * **Answer:** A) Kerberos authentication
11. **What does network segmentation achieve in the context of a Hadoop cluster?**
    * A) Improved data replication speed
    * B) Improved network performance and fault isolation
    * C) Better job scheduling
    * D) Lower hardware costs
    * **Answer:** B) Improved network performance and fault isolation
12. **Which of the following network components ensures high throughput and low latency in Hadoop clusters?**
    * A) High-speed Ethernet network
    * B) Routers
    * C) Modems
    * D) Network printers
    * **Answer:** A) High-speed Ethernet network
13. **Which Hadoop cluster management tool is most commonly used for resource allocation and job scheduling?**
    * A) Ambari
    * B) YARN
    * C) Zookeeper
    * D) HDFS
    * **Answer:** B) YARN
14. **When scaling a Hadoop cluster, what key factor should be considered for balancing the load across all nodes?**
    * A) RAM and CPU specifications
    * B) Number of file system replicas
    * C) Data locality and network bandwidth
    * D) Total number of users
    * **Answer:** C) Data locality and network bandwidth
15. **Which storage method is recommended for handling large amounts of unstructured data in a Hadoop cluster?**
    * A) Object storage
    * B) File storage
    * C) Relational database storage
    * D) NoSQL storage
    * **Answer:** A) Object storage
16. **What is the primary consideration for Hadoop cluster network topology when planning for large-scale deployments?**
    * A) Cluster size and data flow patterns
    * B) Number of cluster administrators
    * C) Availability of backup power
    * D) Cost of electricity
    * **Answer:** A) Cluster size and data flow patterns
17. **Which type of network switch provides optimal performance for large Hadoop clusters?**
    * A) Managed switch with low latency
    * B) Consumer-grade switch
    * C) Unmanaged switch
    * D) Wireless switch
    * **Answer:** A) Managed switch with low latency
18. **In Hadoop clusters, the resource management is handled by which component in Hadoop 2.x?**
    * A) ResourceManager (YARN)
    * B) JobTracker
    * C) NameNode
    * D) DataNode
    * **Answer:** A) ResourceManager (YARN)
19. **When configuring hardware for a Hadoop cluster, what is the best practice for ensuring optimal performance?**
    * A) Use identical hardware components across the cluster
    * B) Use low-cost hardware
    * C) Prioritize network components over storage
    * D) Always use the latest processors
    * **Answer:** A) Use identical hardware components across the cluster
20. \*\*What is the **role of data locality in improving the performance of a Hadoop cluster?**
    * A) It reduces network traffic by keeping data close to the compute nodes
    * B) It increases the load on DataNodes
    * C) It speeds up disk I/O operations
    * D) It reduces the number of replicas
    * **Answer:** A) It reduces network traffic by keeping data close to the compute nodes

Session 13 :

**Cluster Maintenance**

o Managing Hadoop Process both with script and manually

o HDFS Maintenance tasks - Adding, decommissioning data node & etc.

o MapReduce Maintenance tasks - Adding, decommissioning Taskt

o racker, killing job/task & etc.

o Backup & Recovery

### **Easy Questions (15)**

1. **Which of the following is the most common way to manage Hadoop processes?**
   * A) Using the **Hadoop command-line interface**
   * B) Using GUI tools
   * C) By editing configuration files directly
   * D) Through email notifications
   * **Answer:** A) Using the **Hadoop command-line interface**
2. **Which of the following tasks is part of HDFS maintenance?**
   * A) Decommissioning a DataNode
   * B) Adding a new JobTracker
   * C) Scaling MapReduce tasks
   * D) Managing YARN queues
   * **Answer:** A) Decommissioning a DataNode
3. **What is the first step in decommissioning a DataNode in HDFS?**
   * A) Shut down the DataNode
   * B) Run the decommission command in the NameNode
   * C) Modify the cluster configuration
   * D) Move data from the DataNode to a backup server
   * **Answer:** B) Run the decommission command in the NameNode
4. **To manage a MapReduce job, which of the following commands is used?**
   * A) hadoop fs -put
   * B) hadoop job -kill
   * C) hadoop job -status
   * D) hadoop fs -ls
   * **Answer:** B) hadoop job -kill
5. **Which of the following is typically used to start/stop Hadoop services?**
   * A) Hadoop GUI
   * B) start-all.sh and stop-all.sh
   * C) SQL queries
   * D) Configuration files
   * **Answer:** B) start-all.sh and stop-all.sh
6. **What is the primary tool for managing HDFS in Hadoop?**
   * A) Hadoop JobTracker
   * B) Hadoop Distributed FileSystem (HDFS) shell
   * C) HDFS Management Console
   * D) Hadoop MapReduce interface
   * **Answer:** B) Hadoop Distributed FileSystem (HDFS) shell
7. **Which of the following is an important task when adding a DataNode to a Hadoop cluster?**
   * A) Reboot the NameNode
   * B) Update the hdfs-site.xml file on all nodes
   * C) Format the new DataNode
   * D) Move data from old DataNodes to the new one
   * **Answer:** B) Update the hdfs-site.xml file on all nodes
8. **What does decommissioning a DataNode do in Hadoop?**
   * A) It removes the DataNode from the cluster and redistributes its data
   * B) It adds the DataNode to a maintenance list
   * C) It restarts the DataNode
   * D) It increases the replication factor of its data
   * **Answer:** A) It removes the DataNode from the cluster and redistributes its data
9. **Which command is used to monitor Hadoop jobs?**
   * A) hadoop job -status
   * B) hadoop dfs -status
   * C) hadoop job -kill
   * D) hadoop fs -status
   * **Answer:** A) hadoop job -status
10. **Which file stores the configuration for HDFS?**
    * A) core-site.xml
    * B) hdfs-site.xml
    * C) mapred-site.xml
    * D) yarn-site.xml
    * **Answer:** B) hdfs-site.xml
11. **What is the purpose of the hadoop fs -du command?**
    * A) To display file size information in HDFS
    * B) To list files in HDFS
    * C) To check disk usage on the local filesystem
    * D) To delete files from HDFS
    * **Answer:** A) To display file size information in HDFS
12. **Which Hadoop service is responsible for tracking jobs and monitoring their status?**
    * A) ResourceManager
    * B) JobTracker
    * C) NameNode
    * D) DataNode
    * **Answer:** B) JobTracker
13. **What does the hadoop job -kill command do?**
    * A) Starts a MapReduce job
    * B) Displays the status of a running job
    * C) Terminates a running MapReduce job
    * D) Restarts the ResourceManager
    * **Answer:** C) Terminates a running MapReduce job
14. **Which of the following commands is used to manage MapReduce tasks?**
    * A) hadoop fs -get
    * B) hadoop job -list
    * C) hadoop job -status
    * D) hadoop mapred -status
    * **Answer:** C) hadoop job -status
15. \*\*Which of the following is a **critical step in Hadoop cluster backup?**
    * A) Backing up the HDFS metadata
    * B) Rebuilding the ResourceManager
    * C) Stopping all jobs
    * D) Decommissioning all DataNodes
    * **Answer:** A) Backing up the HDFS metadata

### **Intermediate Questions (15)**

1. **Which of the following tasks is essential for decommissioning a TaskTracker in Hadoop?**
   * A) Shut down the TaskTracker
   * B) Remove the TaskTracker’s configuration from the master node
   * C) Stop the JobTracker
   * D) Manually move tasks from the decommissioned node
   * **Answer:** B) Remove the TaskTracker’s configuration from the master node
2. **What is the default replication factor in HDFS when a DataNode is decommissioned?**
   * A) 2
   * B) 3
   * C) 5
   * D) 1
   * **Answer:** B) 3
3. **When adding a new DataNode to HDFS, what must be updated to make it part of the cluster?**
   * A) hdfs-site.xml file
   * B) core-site.xml file
   * C) yarn-site.xml file
   * D) ResourceManager configuration
   * **Answer:** A) hdfs-site.xml file
4. \*\*Which of the following is **not a standard Hadoop maintenance task for HDFS?**
   * A) Decommissioning a DataNode
   * B) Adding a DataNode
   * C) Monitoring disk usage
   * D) Updating MapReduce configurations
   * **Answer:** D) Updating MapReduce configurations
5. **What is the purpose of using balancer in HDFS?**
   * A) To balance the load across all DataNodes
   * B) To create backup replicas
   * C) To replicate missing files
   * D) To optimize MapReduce task distribution
   * **Answer:** A) To balance the load across all DataNodes
6. \*\*What is the **best practice for backup and recovery in a Hadoop cluster?**
   * A) Regularly backup the entire HDFS data directory
   * B) Only backup the HDFS metadata
   * C) Backup only critical job logs
   * D) Do not perform backups in a Hadoop cluster
   * **Answer:** A) Regularly backup the entire HDFS data directory
7. **Which of the following is a critical maintenance task for MapReduce jobs?**
   * A) Kill running jobs when they are stuck
   * B) Backup all job logs
   * C) Restart the DataNode frequently
   * D) Update MapReduce scripts regularly
   * **Answer:** A) Kill running jobs when they are stuck
8. **Which command is used to view the status of HDFS data replication?**
   * A) hadoop fs -ls
   * B) hadoop dfsadmin -report
   * C) hadoop fs -du
   * D) hadoop dfs -status
   * **Answer:** B) hadoop dfsadmin -report
9. **How is MapReduce job recovery typically handled when a job fails?**
   * A) Restart the JobTracker
   * B) Re-run the job from the beginning
   * C) Use the ResourceManager to recover job states
   * D) Automatically recover from checkpoint files
   * **Answer:** C) Use the ResourceManager to recover job states
10. **What happens when a DataNode fails in HDFS?**
    * A) DataNode is removed from the cluster and its data is replicated
    * B) The NameNode will stop all operations
    * C) Data is deleted from the HDFS
    * D) The job will fail immediately
    * **Answer:** A) DataNode is removed from the cluster and its data is replicated
11. **Which of the following is not required for decommissioning a DataNode in Hadoop?**
    * A) Remove the DataNode’s host from the configuration files
    * B) Ensure data is replicated before decommissioning
    * C) Stop the HDFS services
    * D) Move data to an external storage
    * **Answer:** D) Move data to an external storage
12. **Which of the following is a task involved in maintaining MapReduce jobs?**
    * A) Setting input and output paths for the job
    * B) Re-running failed jobs
    * C) Changing YARN configurations
    * D) Adding new TaskTrackers
    * **Answer:** B) Re-running failed jobs
13. **What should you do first when planning a Hadoop cluster backup?**
    * A) Backup the cluster configuration files
    * B) Backup the HDFS metadata
    * C) Backup all the data stored in HDFS
    * D) Verify the cluster’s operational status
    * **Answer:** B) Backup the HDFS metadata
14. **What is the use of the hadoop fs -rm command in Hadoop?**
    * A) To remove a directory from the HDFS
    * B) To remove a file from local file system
    * C) To remove a MapReduce job
    * D) To decommission a DataNode
    * **Answer:** A) To remove a directory from the HDFS
15. **What is one of the main tasks in Hadoop cluster maintenance regarding TaskTrackers?**
    * A) Add or remove TaskTrackers based on job distribution
    * B) Monitor MapReduce job logs
    * C) Add new nodes for TaskTracker redundancy
    * D) Update configuration files for TaskTrackers
    * **Answer:** A) Add or remove TaskTrackers based on job distribution

### **Hard Questions (20)**

1. **What happens to MapReduce jobs if a TaskTracker is decommissioned or fails?**
   * A) Jobs are immediately reassigned to another available TaskTracker
   * B) The JobTracker will retry the task on the failed TaskTracker
   * C) Jobs get stuck and need to be restarted manually
   * D) Job completion is delayed indefinitely
   * **Answer:** A) Jobs are immediately reassigned to another available TaskTracker
2. **In a Hadoop cluster, what is the purpose of "balancing" HDFS data?**
   * A) To redistribute blocks evenly across all DataNodes for performance
   * B) To recover lost files from HDFS
   * C) To create additional copies of critical files
   * D) To compress data on all DataNodes
   * **Answer:** A) To redistribute blocks evenly across all DataNodes for performance
3. \*\*Which of the following **configuration files** is **critical for Hadoop cluster backup and recovery?**
   * A) hdfs-site.xml
   * B) mapred-site.xml
   * C) core-site.xml
   * D) yarn-site.xml
   * **Answer:** A) hdfs-site.xml
4. \*\*What is the **impact of a failed NameNode on the Hadoop cluster?**
   * A) Entire cluster becomes inaccessible
   * B) Jobs continue to run without issues
   * C) DataNode replication will fail
   * D) Only jobs in the queue will be lost
   * **Answer:** A) Entire cluster becomes inaccessible
5. **In terms of cluster management, what does kerberos authentication provide?**
   * A) Secure authentication for users accessing Hadoop resources
   * B) Faster job execution
   * C) Improved data compression
   * D) More efficient HDFS replication
   * **Answer:** A) Secure authentication for users accessing Hadoop resources
6. **What is a critical challenge when adding a new node to a large Hadoop cluster?**
   * A) Balancing the load across nodes
   * B) Upgrading the Hadoop version
   * C) Finding space for the new node
   * D) Reducing job execution time
   * **Answer:** A) Balancing the load across nodes
7. **How do you ensure high availability for the HDFS NameNode?**
   * A) Deploy a secondary NameNode
   * B) Use a backup NameNode in case of failure
   * C) Replicate NameNode data across multiple DataNodes
   * D) Deploy the NameNode on a single node
   * **Answer:** B) Use a backup NameNode in case of failure
8. **What is the primary purpose of the Hadoop balancer tool?**
   * A) To manage disk space and move HDFS blocks for even distribution
   * B) To encrypt sensitive data
   * C) To monitor MapReduce job execution time
   * D) To manage cluster hardware failures
   * **Answer:** A) To manage disk space and move HDFS blocks for even distribution
9. **What type of storage configuration is most effective for Hadoop backup and recovery?**
   * A) Redundant Array of Independent Disks (RAID)
   * B) Local disk storage
   * C) Cloud storage
   * D) Tape storage
   * **Answer:** A) Redundant Array of Independent Disks (RAID)
10. **In a large Hadoop cluster, how can you effectively monitor and maintain performance?**
    * A) By using monitoring tools like Ambari or Ganglia
    * B) Manually tracking resource usage via logs
    * C) Through email alerts for performance issues
    * D) By upgrading nodes regularly
    * **Answer:** A) By using monitoring tools like Ambari or Ganglia
11. **What type of maintenance task should be done when moving large amounts of data across DataNodes in a Hadoop cluster?**
    * A) Run a disk balancer to even out data across nodes
    * B) Restart the cluster
    * C) Perform a cluster-wide reboot
    * D) Compress the data during movement
    * **Answer:** A) Run a disk balancer to even out data across nodes
12. \*\*What is the **role of TaskTracker in Hadoop MapReduce jobs?**
    * A) Assigning jobs to workers and monitoring execution
    * B) Collecting results from jobs and writing to HDFS
    * C) Running the actual tasks and sending updates to JobTracker
    * D) Managing replication factors for MapReduce data
    * **Answer:** C) Running the actual tasks and sending updates to JobTracker
13. **What should you do before performing a backup of a Hadoop cluster?**
    * A) Ensure that all jobs are completed or stopped
    * B) Decommission all DataNodes
    * C) Encrypt all HDFS data
    * D) Upgrade the Hadoop version
    * **Answer:** A) Ensure that all jobs are completed or stopped
14. \*\*What is **one common problem with backup in large Hadoop clusters?**
    * A) Slow data recovery speeds
    * B) Inconsistent backup of HDFS metadata
    * C) Overuse of resources
    * D) Too much job interference during backup
    * **Answer:** B) Inconsistent backup of HDFS metadata
15. **Which of the following is a key challenge in maintaining Hadoop on large clusters?**
    * A) Balancing and optimizing hardware resources
    * B) Ensuring 100% uptime
    * C) Fixing the network during job failures
    * D) Synchronizing all node clocks
    * **Answer:** A) Balancing and optimizing hardware resources
16. **In the event of a DataNode failure, what is the first action to take in maintaining HDFS availability?**
    * A) Replicate data blocks to other DataNodes
    * B) Stop the NameNode
    * C) Remove the failed DataNode from the cluster
    * D) Wait for automatic replication to occur
    * **Answer:** A) Replicate data blocks to other DataNodes
17. **Which of the following is a critical maintenance task for MapReduce jobs?**
    * A) Adjusting resource allocation for job types
    * B) Backing up completed jobs
    * C) Cleaning up HDFS after job completion
    * D) Monitoring job status frequently
    * **Answer:** A) Adjusting resource allocation for job types
18. **What is the main benefit of TaskTracker decommissioning?**
    * A) Reassign jobs to other TaskTrackers for better load balancing
    * B) Prevent jobs from running on the decommissioned node
    * C) Remove jobs from the queue
    * D) Prevent the TaskTracker from accepting any new tasks
    * **Answer:** B) Prevent jobs from running on the decommissioned node
19. \*\*How does **HDFS ensure data consistency during recovery?**
    * A) By replicating lost blocks automatically
    * B) By using read-only copies of blocks
    * C) By applying a checkpoint mechanism
    * D) By halting all jobs until recovery is complete
    * **Answer:** A) By replicating lost blocks automatically
20. **Which of the following should be regularly monitored in a Hadoop cluster for maintenance purposes?**
    * A) Disk health, CPU usage, and network throughput
    * B) The status of individual jobs
    * C) Only the availability of DataNodes
    * D) Only job execution time
    * **Answer:** A) Disk health, CPU usage, and network throughput