1. What is Hadoop?

a) A programming language

b) A distributed storage system

c) A relational database management system

d) A data visualization tool

ans- b) A distributed storage system

2. Hadoop is written in which programming language?

a) Java

b) Python

c) C++

d) Ruby

ans- Java

3. The core components of Hadoop are:

a) HDFS and YARN

b) MapReduce and YARN

c) HDFS and MapReduce

d) HDFS, MapReduce, and YARN

ans- d) HDFS, MapReduce, and YARN

4. What does HDFS stand for?

a) Hadoop Distributed File System

b) Hadoop Data File Storage

c) High-Density File System

d) High-Definition File Storage

ans - a) Hadoop Distributed File System

5. Which component of Hadoop is responsible for resource management and job scheduling?

a) HDFS

b) MapReduce

c) YARN

d) HBase

ans - c) YARN

6. What is the primary function of MapReduce in Hadoop?

a) Data storage and retrieval

b) Data visualization

c) Data processing and computation

d) Data security and encryption

ans - c) Data processing and computation

7. What is the default storage format in Hadoop MapReduce?

a) XML

b) CSV

c) JSON

d) SequenceFile

ans - d) SequenceFile

8. Which Hadoop ecosystem component is used for real-time data processing and analytics?

a) Hive

b) Pig

c) HBase

d) Oozie

ans - c) HBase

9. Which programming paradigm does Hadoop MapReduce follow?

a) Object-oriented programming (OOP)

b) Procedural programming

c) Functional programming

d) Event-driven programming

ans - c) Functional programming

10. In Hadoop, what is a NameNode?

a) A node responsible for storing actual data in HDFS

b) A node that executes MapReduce tasks

c) A node responsible for managing metadata in HDFS

d) A node used for running YARN applications

Answers: c) A node responsible for managing metadata in HDFS

11. Which component of Hadoop is responsible for storing data in a distributed manner?

a) Hadoop Distributed File System (HDFS)

b) Hadoop Query Language (HQL)

c) Hadoop Application Manager (HAM)

d) Hadoop Job Tracker (HJT)

Ans - a) Hadoop Distributed File System (HDFS)

12.The core processing unit in Hadoop is called:

a) MapTask

b) ReduceTask

c) JobTask

d) CoreTask

Ans - a) MapTask

13.What is the role of the MapReduce framework in Hadoop?

a) It provides real-time data processing capabilities.

b) It allows users to interact with Hadoop through a web interface.

c) It enables the distributed processing of large datasets.

d) It is responsible for managing data replication in HDFS.

Ans - c) It enables the distributed processing of large datasets.

14. What does the term "NameNode" refer to in Hadoop?

a) The node responsible for executing Map tasks.

b) The node responsible for executing Reduce tasks.

c) The node that manages the metadata of files in HDFS.

d) The node that handles the web interface for Hadoop administration.

Ans - c) The node that manages the metadata of files in HDFS.

15. What is the default replication factor in HDFS?

a) 1

b) 2

c) 3

d) 4

Ans - c) 3

16. Which Apache project provides a high-level abstraction for creating complex data workflows in Hadoop?

a) Apache Pig

b) Apache Hive

c) Apache Oozie

d) Apache Flume

Ans - c) Apache Oozie

17. What is the main advantage of Hadoop's distributed storage and processing capabilities?

a) Low-latency real-time data processing

b) High availability of data without replication

c) Scalability and ability to handle big data

d) Easy integration with relational databases

Answers: c) Scalability and ability to handle ability to handle big data

18. The core components of Hadoop include:

a) Hadoop Distributed File System (HDFS) and Hadoop MapReduce

b) Hadoop Query Language (HQL) and Hadoop Streaming

c) Hadoop SQL and HadoopDB

d) Hadoop Storage System (HSS) and Hadoop Compute Engine (HCE)

Ans - a) Hadoop Distributed File System (HDFS) and Hadoop MapReduce

19. In Hadoop MapReduce, what does the Mapper do?

a) It processes input data and produces key-value pairs as output.

b) It aggregates the output from multiple Mappers.

c) It sorts the data before passing it to the Reducer.

d) It distributes the data across different nodes in the cluster.

Ans - a) It processes input data and produces key-value pairs as output.

20. In Hadoop, what does the term "Data Node" refer to?

a) A node that stores metadata about the HDFS

b) A node that runs the MapReduce jobs

c) A node that stores and manages the actual data in HDFS

d) A node that acts as a master in the Hadoop cluster

Ans - c) A node that stores and manages the actual data in HDFS

21. The secondary Name Node in Hadoop is responsible for:

a) Maintaining a backup of the HDFS metadata

b) Storing multiple replicas of data blocks for fault tolerance

c) Distributing data across multiple nodes for parallel processing

d) Managing the resources and scheduling of MapReduce jobs

Ans - a) Maintaining a backup of the HDFS metadata

22. Which Hadoop component is responsible for managing resources and job scheduling in a cluster?

a) ResourceManager

b) DataNode

c) NameNode

d) TaskTracker

Ans a) ResourceManager

23. Which of the following is NOT a V's of Big Data associated with Hadoop?

a) Volume

b) Velocity

c) Validation

d) Variety

Ans - c) Validation

24. Which Hadoop ecosystem tool is used for real-time stream processing of data?

a) Apache Hive

b) Apache Pig

c) Apache HBase

d) Apache Kafka

Ans - d) Apache Kafka

25. In Hadoop, what is a "Mapper"?

a) A node responsible for executing user-defined MapReduce functions

b) A daemon that handles data storage in HDFS

c) A tool for querying and managing data in Hadoop

d) A component that manages resource allocation in the cluster

Ans - a) A node responsible for executing user-defined MapReduce functions

26. What is the primary purpose of HBase in the Hadoop ecosystem?

a) To process and analyze large datasets in parallel

b) To provide real-time querying of Hadoop data

c) To manage job scheduling in Hadoop clusters

d) To provide distributed storage for Hadoop data

Answers: b) To provide real-time querying of Hadoop data

27. Hadoop is designed to handle which type of data?

a) Structured data

b) Unstructured data

c) Semi-structured data

d) All of the above

Ans - d) All of the above

28. HDFS is:

a) A database management system

b) A file system for storing large volumes of data

c) A distributed data processing engine

d) A distributed caching system

Ans - b) A file system for storing large volumes of data

29. The processing model used in Hadoop is:

a) MapReduce

b) Spark

c) Hive

d) Pig

Ans - MapReduce

30. What does YARN stand for in the Hadoop ecosystem?

a) Yet Another Resource Negotiator

b) Yet Another Resource Navigator

c) Yet Another Resource Node

d) Yet Another Resource NameNode

Ans - a) Yet Another Resource Negotiator

31. Which Hadoop component is responsible for distributing tasks to nodes in the cluster?

a) JobTracker

b) NameNode

c) ResourceManager

d) TaskTracker

Ans - c) ResourceManager

32. Which Hadoop ecosystem tool is used for SQL-based querying and analysis of data?

a) HBase

b) Hive

c) ZooKeeper

d) Oozie

Ans - b) Hive

33. What is the default scheduler in Hadoop YARN?

a) CapacityScheduler

b) FairScheduler

c) FifoScheduler

d) ResourceScheduler

Ans - c) FifoScheduler

34. Which Hadoop ecosystem component is used for real-time data processing?

a) HBase

b) Pig

c) Hive

d) Spark

Ans - d) Spark

35. What is the primary advantage of Hadoop's distributed computing model?

a) Faster data processing

b) Lower hardware costs

c) Better data security

d) Fault tolerance and scalability

Ans - d) Fault tolerance and scalability

36. What is the primary function of Hadoop MapReduce?

a) Data storage

b) Data visualization

c) Data processing

d) Data security

Ans - c) Data processing

37. What is the purpose of DataNodes in HDFS?

a) Storing the metadata of files and directories

b) Running MapReduce tasks on data

c) Storing and managing the actual data blocks of files

d) Providing the Hadoop cluster's computing resources

Ans - c) Storing and managing the actual data blocks of files

38. In Hadoop MapReduce, what is the function of the Mapper?

a) Sorting and shuffling the intermediate key-value pairs

b) Reducing the data into a smaller set of key-value pairs

c) Reading data from HDFS and generating key-value pairs

d) Coordinating the execution of the MapReduce job

Ans - c) Reading data from HDFS and generating key-value pairs

39. What is the primary responsibility of the Reducer in Hadoop MapReduce?

a) Writing data back to HDFS

b) Handling data input/output operations

c) Aggregating and processing the data produced by Mappers

d) Distributing the MapReduce tasks across the cluster

Ans - c) Aggregating and processing the data produced by Mappers

40. Which component of Hadoop is responsible for resource management and job scheduling?

a) NameNode

b) DataNode

c) ResourceManager

d) TaskTracker

Ans - c) ResourceManager

41. In Hadoop, which service ensures high availability and fault tolerance for the NameNode?

a) Secondary NameNode

b) ResourceManager

c) DataNode

d) JobTracker

Ans - a) Secondary NameNode

42. How does Hadoop ensure data fault tolerance in HDFS?

a) By replicating data blocks across multiple DataNodes

b) By creating backups on external storage systems

c) By compressing data to reduce the risk of data loss

d) By automatically moving data to different nodes periodically

Ans - a) By replicating data blocks across multiple DataNodes

43. Which statement is true about Hadoop?

a) It is a traditional relational database management system.

b) It can only process structured data formats.

c) It is a framework for processing and analyzing big data in a distributed environment.

d) It is mainly used for real-time data processing.

Ans - c) It **is a framework for processing and analyzing big data in a distributed environment.**

YARN

1. What is YARN in the Hadoop ecosystem?

a) A data processing engine

b) A cluster management technology

c) A NoSQL database

d) A machine learning framework

Ans - b) A cluster management technology

2. What is the primary function of the YARN Resource Manager?

a) Managing the HDFS (Hadoop Distributed File System)

b) Scheduling jobs on the cluster

c) Managing data storage in Hadoop

d) Performing data processing tasks

Ans - b) Scheduling jobs on the cluster

3. What is a Node Manager in YARN responsible for?

a) Allocating resources to applications

b) Storing data in Hadoop

c) Monitoring the health of the Hadoop cluster

d) Managing the Name Node in HDFS

Ans - c) Monitoring the health of the Hadoop cluster

4. Which YARN component is responsible for restarting containers in case of failures?

a) Application Master

b) Resource Manager

c) Node Manager

d) Data Node

Ans - c) Node Manager

5. How does YARN handle resource management for different applications running on the cluster?

a) Each application gets its dedicated cluster with separate nodes

b) YARN dynamically allocates resources to each application based on its requirements

c) Resource allocation is fixed, and applications must share resources

d) YARN randomly assigns resources to applications

Ans - b) YARN dynamically allocates resources to each application based on its requirements

6. Which file in YARN specifies the resources required and the application's execution command?

a) core-site.xml

b) hdfs-site.xml

c) yarn-site.xml

d) application.xml

Ans - d) application.xml

7. What does the YARN Capacity Scheduler ensure?

a) Fair allocation of resources to all applications

b) Immediate termination of low-priority applications

c) Guaranteed capacity for each application in the cluster

d) Automatic scaling of the cluster based on application demand

Ans - c) Guaranteed capacity for each application in the cluster

8. What is the role of the Application Master in YARN?

a) Managing resources for the entire cluster

b) Tracking the progress of MapReduce jobs

c) Monitoring the health of the HDFS

d) Coordinating and managing the lifecycle of an application on the cluster

Ans - d) Coordinating and managing the lifecycle of an application on the cluster

9. Which command is used to submit a YARN application to the Resource Manager?

a) Hadoop dfs

b) yarn jar

c) mapred submit

d) yarn app

Ans - b) yarn jar

10. What happens if a YARN application's resource requirements exceed the available cluster capacity?

a) The application will be paused until resources are available.

b) YARN will automatically scale the cluster to accommodate the application.

c) The application will be terminated with an error.

d) The application will run with reduced resources.

Answers: a) The application will be paused until resources are available.

11. What is YARN in Apache Hadoop?

a) Yet Another Reliable Network

b) Yet Another Resource Navigator

c) Yet Another Resource Negotiator

d) Yet Another Replication Node

Ans - c) Yet Another Resource Negotiator

12. Which component in YARN is responsible for resource allocation to applications?

a) ResourceManager

b) NodeManager

c) Application Master

d) DataNode

Ans - a) ResourceManager

13. What is the primary function of the YARN ResourceManager?

a) Manage data storage on Hadoop Distributed File System (HDFS)

b) Execute application tasks on the cluster

c) Manage and allocate cluster resources to applications

d) Monitor the health of individual nodes in the cluster

Ans - c) Manage and allocate cluster resources to applications

14. In YARN, what does the Application Master do?

a) Allocates resources to different applications

b) Manages resources on the DataNodes

c) Executes tasks for a specific application and negotiates resource containers

d) Acts as a master node for the entire YARN cluster

Ans - c) Executes tasks for a specific application and negotiates resource containers

15. Which file specifies the resources (CPU, memory, etc.) required by a YARN application?

a) yarn-site.xml

b) hdfs-site.xml

c) core-site.xml

d) yarn.xml

Ans - d) yarn.xml

16. What are the two main components of a YARN application?

a) JobTracker and TaskTracker

b) Resource Manager and NodeManager

c) Application Master and DataNode

d) Name Node and ResourceManager

Ans - b) Resource Manager and NodeManager

17. Which command is used to submit a YARN application?

a) hadoop jar

b) yarn submit

c) yarn application -submit

d) yarn jar

Ans - d) yarn jar

18. Which YARN daemon is responsible for managing the execution of individual tasks on a node?

a) NodeManager

b) ResourceManager

c) DataNode

d) Application Master

Ans - a) NodeManager

19. In a YARN cluster, how does the Node Managerobtain resources to run tasks?

a) It requests resources from the ResourceManager.

b) It takes resources from the Node Managerof another cluster.

c) It allocates resources based on its own configuration.

d) It requests resources directly from the HDFS.

Ans - a) It requests resources from the ResourceManager.

20. What happens if the Application Master fails during the execution of a YARN application?

a) The Resource Manager restarts the Application Master on a different node.

b) The entire YARN cluster restarts.

c) The affected application is automatically terminated.

d) The Node Managertakes over the role of the Application Master .

Answers: a) The Resource Manager restarts the Application Master on a different node.

21. What is the primary role of YARN in Hadoop?

a) Data storage

b) Data processing

c) Resource management

d) Network communication

Ans - c) Resource management

22. In YARN, what is the function of the Application Master?

a) Manage the execution of tasks within an application

b) Manage resources in the cluster

c) Coordinate communication between nodes in the cluster

d) Allocate resources to applications

Ans - a) Manage the execution of tasks within an application

23. What does a Resource Manager do when it receives a resource request from an Application Master?

a) Allocate resources to the requesting application

b) Deny the request if the cluster is already at full capacity

c) Delay the request until resources become available

d) Ask the Node Managerto allocate resources directly

Ans - a) Allocate resources to the requesting application

24. How does YARN handle the failures of NodeManagers?

a) It reassigns the failed tasks to other nodes

b) It stops the entire application and notifies the ResourceManager

c) It restarts the Node Managerautomatically

d) It waits for the Node Manager to recover on its own

Ans - a) It reassigns the failed tasks to other nodes

25. In YARN, what are containers?

a) Virtual machines used to isolate applications

b) Data structures used to store intermediate results

c) Units of resource allocation (CPU and RAM) for running application tasks

d) Communication channels between Node Managers and the Resource Manager

Ans - c) Units of resource allocation (CPU and RAM) for running application tasks

26. What is the function of the YARN Resource Manager's scheduler?

a) Ensure data is stored redundantly across the cluster

b) Determine the order in which applications are executed

c) Monitor the health of nodes in the cluster

d) Allocate and manage resources for running applications

Ans - d) Allocate and manage resources for running applications

27. Which of the following components runs and monitors the tasks of a specific application in YARN?

a) ResourceManager

b) NodeManager

c) Application Master

d) HDFS NameNode

Ans - c) Application Master

28. What is the primary goal of YARN's resource management?

a) Ensure security and encryption of data in Hadoop

b) Optimize storage of data on HDFS

c) Efficiently utilize cluster resources and allocate them to applications

d) Minimize the number of nodes in the Hadoop cluster

Ans - c) Efficiently utilize cluster resources and allocate them to applications

29. Which configuration file is used to adjust the YARN Resource Manager settings in Hadoop?

a) hdfs-site.xml

b) yarn-site.xml

c) mapred-site.xml

d) core-site.xml

Ans - b) yarn-site.xml

30. In YARN, which component is responsible for tracking the status of individual tasks and coordinating their execution?

a) ResourceManager

b) Application Master

c) NodeManager

d) JobTracker

Ans - b) Application Master

31. How does YARN support multi-tenancy in Hadoop?

a) It allows multiple Hadoop clusters to run on the same hardware.

b) It allows multiple users to share the same YARN cluster while providing isolation.

c) It enables YARN to run on multiple operating systems simultaneously.

d) It allows YARN to distribute data across multiple nodes.

Ans - b) It allows multiple users to share the same YARN cluster while providing isolation.

32. Which of the following is NOT a scheduling policy supported by YARN?

a) Capacity Scheduler

b) Fair Scheduler

c) FIFO Scheduler

d) Round-Robin Scheduler

Ans – Round-robin Scheduler

33. What happens if a Node Manager fails during the execution of an application in YARN?

a) The application fails, and all progress is lost.

b) The Resource Manager automatically recovers the failed Node Manager.

c) The Application Master restarts the failed tasks on other healthy nodes.

d) The entire YARN cluster becomes unavailable until the Node Manager is fixed.

Ans - c) The Application Master restarts the failed tasks on other healthy nodes.

34. In YARN, what is the role of the Application Master?

a) Manage resources on all the nodes in the cluster.

b) Manage resources for a specific application and negotiate resources with the Resource Manager.

c) Act as the primary Name Node for the Hadoop cluster.

d) Coordinate the distribution of input data to the worker nodes.

Ans - b) Manage resources for a specific application and negotiate resources with the Resource Manager.

35. Which web UI provides information about the overall cluster resource utilization and applications running in YARN?

a) Resource Manager UI

b) Node ManagerUI

c) Application Master UI

d) HDFS Name Node UI

Ans - a) Resource Manager UI

36. Which command is used to submit a YARN application to the ResourceManager?

a) yarn run

b) hadoop submit

c) mapred submit

d) yarn application-submit

Ans - d) yarn application-submit

37. Which component in YARN is responsible for managing resources and scheduling tasks?

a) ResourceManager

b) NameNode

c) NodeManager

d) DataNode

Ans - a) ResourceManager

38. In YARN, what is the function of the ApplicationMaster?

a) It manages the resources on the cluster.

b) It processes and analyzes data using MapReduce.

c) It handles the lifecycle of a specific application on the cluster.

d) It stores and retrieves data from HDFS.

Ans - c) It handles the lifecycle of a specific application on the cluster.

39. What does the term "container" refer to in the context of YARN?

a) A physical machine in the cluster

b) A virtual machine running on the cluster

c) An isolated execution environment for an application's process

d) A storage location for Hadoop data

Ans - c) An isolated execution environment for an application's process

40. Which YARN component is responsible for maintaining information about available cluster resources?

a) NodeManager

b) ApplicationMaster

c) ResourceManager

d) DataNode

Ans - c) ResourceManager

41. What is the role of the NodeManager in YARN?

a) It manages the execution of tasks for a specific application.

b) It manages the distribution of data across the cluster.

c) It monitors the health of the ResourceManager.

d) It monitors the resources and runs on each node in the cluster.

Ans - d) It monitors the resources and runs on each node in the cluster.

42. Which command is used to submit a YARN application to the ResourceManager?

a) hadoop fs -put

b) yarn app -submit

c) yarn jar

d) hadoop jar

Ans - c) yarn jar

43. Which scheduling policy in YARN ensures that each application gets a fair share of cluster resources over time?

a) Capacity Scheduler

b) Fair Scheduler

c) FIFO Scheduler

d) Priority Scheduler

Ans - b) Fair Scheduler

44. In YARN, what is the function of the ResourceManager?

a) It runs the application-specific tasks.

b) It stores and manages the Hadoop data.

c) It handles resource allocation and scheduling across the cluster.

d) It manages the networking and communication between nodes.

Ans - c) It handles resource allocation and scheduling across the cluster.

45. Which service in YARN is responsible for tracking the status and progress of applications?

a) ApplicationTracker

b) ApplicationHistoryServer

c) ApplicationManager

d) ApplicationStatusMonitor

Ans - b) ApplicationHistoryServer

HDFS

What is the primary purpose of HDFS in the Hadoop ecosystem?

a) Real-time data processing

b) Streamlining data pipelines

c) Storing and managing large-scale data

d) Running machine learning algorithms

Ans - c) Storing and managing large-scale data

Which component of Hadoop is responsible for managing storage and replication of data across nodes in the cluster?

a) ResourceManager

b) NameNode

c) DataNode

d) JobTracker

Ans - c) DataNode

What happens if the NameNode in an HDFS cluster fails?

a) The cluster continues to work normally without any issues.

b) The entire HDFS cluster becomes unavailable until the NameNode is restored.

c) DataNodes take over the responsibility of the NameNode until it is restored.

d) The HDFS cluster becomes read-only until the NameNode is restored.

Ans - b) The entire HDFS cluster becomes unavailable until the NameNode is restored.

What is the typical block size used in HDFS?

a) 64 MB

b) 128 MB

c) 256 MB

d) 512 MB

Ans - b) 128 MB

Which HDFS command is used to copy files from the local file system to the HDFS file system?

a) get

b) copyFromLocal

c) put

d) copyToLocal

Ans - c) put

What is the purpose of the Secondary NameNode in HDFS?

a) It acts as a backup for the primary NameNode.

b) It assists the primary NameNode in handling client requests.

c) It stores multiple copies of data for fault tolerance.

d) It helps in data replication across the cluster.

Ans - a) It acts as a backup for the primary NameNode.

In HDFS, what is the default replication factor?

a) 1

b) 2

c) 3

d) 4

Ans - c) 3

Which process is responsible for managing resources and scheduling jobs in Hadoop MapReduce?

a) NameNode

b) ResourceManager

c) TaskTracker

d) JobTracker

Ans - b) ResourceManager

What happens if a DataNode in the HDFS cluster fails?

a) The DataNode is automatically restarted by HDFS.

b) The NameNode replicates the lost data blocks to another DataNode.

c) The data hosted by the failed DataNode becomes unavailable until the node is restored.

d) The remaining DataNodes collectively rebuild the failed DataNode's data.

Ans - c) The data hosted by the failed DataNode becomes unavailable until the node is restored.

What is the secondary storage mechanism used by HDFS for maintaining metadata and namespace information in case of NameNode failure?

a) HBase

b) ZooKeeper

c) YARN

d) HCatalog

Ans - b) ZooKeeper

In HDFS, what is the purpose of the "hadoop fsck" command?

a) To create a new HDFS directory

b) To start the HDFS cluster

c) To check the overall health of the HDFS file system

d) To run a MapReduce job

Ans - c) To check the overall health of the HDFS file system

Which component of Hadoop YARN manages the allocation of resources and tracks job progress?

a) NodeManager

b) ApplicationMaster

c) ResourceManager

d) DataNode

Ans - c) ResourceManager

What happens when a DataNode in HDFS goes offline or fails?

a) The NameNode redistributes its data blocks to other DataNodes in the cluster.

b) The entire HDFS cluster goes offline.

c) The HDFS files stored on the failed DataNode are lost permanently.

d) The ResourceManager takes over the responsibilities of the failed DataNode.

Ans - a) The NameNode redistributes its data blocks to other DataNodes in the cluster.

What is the purpose of the "hadoop fs" command in HDFS?

a) To format the HDFS file system

b) To manage HDFS block size

c) To interact with the HDFS file system by copying, deleting, and listing files

d) To configure HDFS replication factor

Ans - c) To interact with the HDFS file system by copying, deleting, and listing files

HDFS provides fault tolerance by:

a) Storing data on multiple drives within each DataNode

b) Storing data replicas on multiple DataNodes

c) Encrypting the data blocks

d) Using RAID arrays for data storage

Answer: (b) Storing data replicas on multiple DataNodes

HDFS is designed to store:

a) Small files

b) Medium-sized files

c) Large files

d) All file sizes

Answer: (c) Large files

HDFS follows a Master/Slave architecture. The master node is called:

a) DataNode

b) ResourceManager

c) NameNode

d) TaskTracker

Answer: (c) NameNode

What is the role of the DataNodes in HDFS?

a) Store the data blocks of files

b) Manage the metadata of files

c) Execute MapReduce jobs

d) Handle client requests

Answer: (a) Store the data blocks of files

What is the secondary Namenode's role in HDFS?

a) Take over as the primary Namenode if it fails

b) Keep a backup of the data stored in HDFS

c) Assist the primary Namenode in managing metadata

d) Handle read and write requests from clients

Answer: (c) Assist the primary Namenode in managing metadata

Which command is used to copy files from the local file system to HDFS?

a) cpToLocal

b) get

c) copyFromLocal

d) fetch

Answer: (c) copyFromLocal

The command used to list the contents of a directory in HDFS is:

a) ls

b) dir

c) list

d) show

Answer: (a) ls

What is the role of the "fsck" command in HDFS?

a) Initiates a MapReduce job

b) Checks the health of the HDFS cluster

c) Displays the content of a specific file

d) Modifies the replication factor of a file

Ans - b) Checks the health of the HDFS cluster

Which of the following is NOT a characteristic of HDFS?

a) High throughput data access

b) Low latency data access

c) Horizontal scalability

d) Suitable for storing large files

Ans - b) Low latency data access

HDFS uses a block placement policy to decide where to place replicas of a data block. What is the default block placement policy in HDFS?

a) Random placement

b) Round-robin placement

c) Least-used node placement

d) Rack-aware placement

Ans - b) Low latency data access

In HDFS, the process of combining data from multiple DataNodes into a single, coherent result is called:

a) Aggregation

b) Reduction

c) Shuffling

d) MapReduce

Answers: b) Low latency data access

In HDFS, data is stored in:

a) Databases

b) Tables

c) Blocks

d) Partitions

Ans - c) Blocks

What is the main advantage of storing data in smaller blocks in HDFS?

a) Faster data processing

b) Better data compression

c) Improved fault tolerance

d) Reduced memory overhead

Ans - c) Improved fault tolerance

The HDFS NameNode stores:

a) File data

b) File metadata

c) Both file data and metadata

d) Job history of MapReduce tasks

Ans - b) File metadata

What happens if the NameNode in HDFS fails?

a) Data becomes inaccessible until the NameNode is restored

b) Secondary NameNode takes over as the new primary NameNode

c) Data stored in HDFS is unaffected

d) HDFS switches to an alternative distributed file system

Ans - a) Data becomes inaccessible until the NameNode is restored

The process of splitting and distributing data blocks across multiple nodes in a Hadoop cluster is known as:

a) Shuffling

b) Partitioning

c) Replication

d) Data Serialization

Ans - b) Partitioning

Which component in Hadoop is responsible for handling data read and write requests from clients?

a) DataNode

b) NameNode

c) ResourceManager

d) SecondaryNameNode

Ans - a) DataNode

MapReduce

What is MapReduce in Hadoop?

a) A data storage system

b) A query language for Hadoop

c) A programming model for distributed data processing

d) A machine learning library for Hadoop

Ans - c) A programming model for distributed data processing

The primary goal of MapReduce is to:

a) Optimize data storage in Hadoop

b) Process data in real-time

c) Enable distributed data processing on large datasets

d) Provide data encryption for Hadoop clusters

Ans - c) Enable distributed data processing on large datasets

MapReduce works on the principle of:

a) Map and Sort

b) Divide and Conquer

c) Load and Unload

d) Store and Retrieve

Ans - b) Divide and Conquer

Which of the following phases in MapReduce sorts the output of the Map phase before passing it to the Reduce phase?

a) Shuffle

b) Sort

c) Merge

d) Aggregate

Ans - a) Shuffle

In MapReduce, what is the function of the Map task?

a) Aggregating the input data

b) Filtering the output data

c) Processing and transforming input data into key-value pairs

d) Combining the output of multiple Reduce tasks

Ans - c) Processing and transforming input data into key-value pairs

The Reduce task in MapReduce is responsible for:

a) Sorting the output data

b) Combining data from multiple sources

c) Splitting the input data into smaller chunks

Ans - d) Performing the final data processing and generating output

In Hadoop MapReduce, the output of the Mapper is input to the:

a) Reducer

b) Combiner

c) Partitioner

d) Secondary Sort

Ans - a) Reducer

What is the role of the Combiner in MapReduce?

a) Compress the output data before sending it to the Reducer

b) Combine intermediate key-value pairs locally on each Mapper node

c) Shuffle and sort the Mapper output before passing it to the Reducer

d) Perform a secondary sorting of the Mapper output before the Reduce phase

Ans - b) Combine intermediate key-value pairs locally on each Mapper node

Which component is responsible for managing resources and scheduling tasks in Hadoop MapReduce?

a) JobTracker

b) NameNode

c) ResourceManager

d) TaskTracker

Ans - c) ResourceManager

In Hadoop 2.x and later versions, the MapReduce framework has been integrated with which resource management system?

a) YARN (Yet Another Resource Negotiator)

b) Apache ZooKeeper

c) Apache Mesos

d) Apache Hive

Answers: a) YARN (Yet Another Resource Negotiator)

Which of the following phases is responsible for data sorting and shuffling in MapReduce?

a) Map

b) Combine

c) Shuffle and Sort

d) Reduce

Ans - c) Shuffle and Sort

What is the purpose of the Map phase in MapReduce?

a) To perform data sorting

b) To combine data from multiple mappers

c) To process input data and produce key-value pairs

d) To aggregate the final results

Ans - c) To process input data and produce key-value pairs

In MapReduce, what does the Reduce phase do?

a) It aggregates the output of the Map phase.

b) It performs data sorting and shuffling.

c) It processes input data and produces key-value pairs.

d) It distributes data to different mappers.

Ans - a) It aggregates the output of the Map phase.

The MapReduce job is divided into tasks, and each task operates on a portion of the data. What are these tasks called?

a) Sub-jobs

b) Jobs

c) Splits

Ans - d) Tasks

Which component is responsible for managing and monitoring MapReduce jobs in Hadoop?

a) ResourceManager

b) DataNode

c) TaskTracker

Ans - c) TaskTracker

In MapReduce, what is the output key-value pair type from the Map phase to the Reduce phase?

a) Text

b) IntWritable

c) Object

d) NullWritable

Ans - a) Text

Which phase in MapReduce is optional and allows for combining the output of the Map phase before sending it to the Reduce phase?

a) Combine

b) Sort

c) Shuffle

d) Partition

Ans - a) Combine

In a MapReduce job, what is the role of the Combiner function?

a) It combines multiple jobs into one.

b) It reduces the output data size before sending it to the reducers.

c) It defines the input splits for the mappers.

d) It ensures fault tolerance in the MapReduce job.

Answers: b) It reduces the output data size before sending it to the reducers.

What are the two main phases in the MapReduce process?

a) Sorting and Filtering

b) Shuffling and Merging

c) Mapping and Reducing

d) Input and Output

Ans - c) Mapping and Reducing

What is the primary function of the Map phase in MapReduce?

a) Data sorting

b) Data reduction

c) Data transformation

d) Data storage

Ans - c) Data transformation

Which component is responsible for distributing tasks to individual nodes in the Hadoop cluster?

a) ResourceManager

b) DataNode

c) NameNode

d) ApplicationMaster

Ans - a) ResourceManager

In MapReduce, what is the purpose of the Shuffle phase?

a) It performs sorting of the input data.

b) It transfers data between the Map and Reduce phases.

c) It manages the job queue.

d) It compresses the intermediate data.

Ans - b) It transfers data between the Map and Reduce phases.

Which part of the MapReduce framework is responsible for fault tolerance and task recovery?

a) ApplicationMaster

b) ResourceManager

c) DataNode

d) TaskTracker

Ans - a) ApplicationMaster

What is the role of the Reducer in MapReduce?

a) It reads input data and writes the output to HDFS.

b) It processes the data in parallel across multiple nodes.

c) It performs data sorting and shuffling.

d) It combines and aggregates data from the Mapper phase.

Ans - d) It combines and aggregates data from the Mapper phase.

In Hadoop 2.x, what is the purpose of the YARN (Yet Another Resource Negotiator) component?

a) Managing data replication in HDFS

b) Scheduling and managing resources in the cluster

c) Performing data transformation in MapReduce jobs

d) Managing the MapReduce shuffle phase

Ans - b) Scheduling and managing resources in the cluster

What is speculative execution in the context of MapReduce jobs?

a) Running multiple instances of a task on different nodes simultaneously

b) Executing tasks using speculative algorithms to improve performance

c) Re-executing failed tasks on a different node

d) Specifying the execution order of tasks in a job

Ans - a) Running multiple instances of a task on different nodes simultaneously

MapReduce is designed to process which types of data?

a) Structured data only

b) Unstructured data only

c) Semi-structured data only

d) All types of data

Ans - d) All types of data

In the MapReduce paradigm, what does the Map phase do?

a) Aggregates the output data

b) Sorts the input data

c) Processes and filters input data to generate key-value pairs

d) Initiates the Hadoop job

Ans - c) Processes and filters input data to generate key-value pairs

What does the Reduce phase do in MapReduce?

a) Performs a final aggregation on the Map output

b) Processes and filters input data to generate key-value pairs

c) Sorts the output data

d) Initiates the Hadoop job

Ans - a) Performs a final aggregation on the Map output

The intermediate data generated during the Map phase is temporarily stored in:

a) Local disk storage on the TaskTracker node

b) HDFS

c) Amazon S3 (Simple Storage Service)

d) Google Cloud Storage

Ans - a) Local disk storage on the TaskTracker node

Which of the following statements is true about the Mapper and Reducer functions in MapReduce?

a) Both Mapper and Reducer functions are user-defined.

b) Mapper is user-defined, but Reducer is built-in.

c) Reducer is user-defined, but Mapper is built-in.

d) Both Mapper and Reducer functions are built-in

Ans - a) Both Mapper and Reducer functions are user-defined.

What is the primary purpose of the Combiner function in MapReduce?

a) To combine multiple input files into a single file before processing

b) To reduce the output data size before transferring it to the Reducer

c) To compress the input data before processing

d) To create a summary report of the Map phase

Ans - b) To reduce the output data size before transferring it to the Reducer

In a MapReduce job, the number of Reducer tasks is determined by:

a) The number of Map tasks

b) The size of the input data

c) The developer's preference

d) The configuration parameter "mapreduce.reduce.tasks"

Ans - d) The configuration parameter "mapreduce.reduce.tasks"

In Hadoop MapReduce, what is the role of the YARN ResourceManager?

a) Scheduling MapReduce jobs on the cluster

b) Storing intermediate data during the Map phase

c) Managing HDFS metadata

d) Monitoring data nodes in the Hadoop cluster

Ans - a) Scheduling MapReduce jobs on the cluster

In Hadoop MapReduce, where does the output of the job go after the Reduce phase?

a) Local file system of the machine that executed the job

b) HDFS

c) HBase

d) Apache Hive

Ans - b) HDFS

MapReduce is best suited for which type of data processing tasks?

a) Real-time data processing

b) Small-scale data analysis

c) Data processing tasks requiring iterative algorithms

d) Sequential data processing

Ans - c) Data processing tasks requiring iterative algorithms

HIVE

What is Apache Hive?

a) A distributed file system

b) A data integration tool

c) A data warehouse infrastructure built on top of Hadoop

d) A NoSQL database management system

Ans - c) A data warehouse infrastructure built on top of Hadoop

In Hive, data is organized into which structures for storage and processing?

a) Tables

b) Partitions

c) Buckets

d) All of the above

Ans - d) All of the above

Which language is used to write queries in Hive?

a) HiveQL (HQL)

b) SQL

c) Python

d) Java

Ans - a) HiveQL (HQL)

The data manipulation operations in Hive are similar to which standard database language?

a) SQL

b) NoSQL

c) Java

d) HiveQL

Ans - a) SQL

What is Hive's role in Hadoop ecosystem?

a) Hive processes and stores real-time data

b) Hive is used for data ingestion in Hadoop

c) Hive provides a SQL-like interface for data querying and analysis on Hadoop

d) Hive handles resource management in Hadoop clusters

Ans - c) Hive provides a SQL-like interface for data querying and analysis on Hadoop

In Hive, what is the function of a Metastore?

a) It stores metadata information about Hive tables and partitions.

b) It stores the actual data files of Hive tables.

c) It optimizes the Hive query execution plan.

d) It manages the Hadoop cluster's resources.

Ans - a) It stores metadata information about Hive tables and partitions.

Which of the following storage formats are supported by Hive?

a) ORC (Optimized Row Columnar)

b) XML

c) CSV (Comma Separated Values)

d) JSON

e) All of the above

Ans - e) All of the above

What is the role of SerDe in Hive?

a) It stands for "Service and Deployment," managing Hive's configuration and deployment.

b) It is a language used to write UDFs (User-Defined Functions) in Hive.

c) It is responsible for parsing data into and out of Hive tables.

d) It optimizes data storage in Hive tables.

Ans - c) It is responsible for parsing data into and out of Hive tables.

Which of the following can be used to improve Hive query performance?

a) Hive views

b) Indexes

c) Partitions

d) All of the above

Ans - d) All of the above

Hive supports the integration of external systems to perform tasks within Hive queries. What are these integrations called?

a) SerDe

b) HBase

c) UDF (User-Defined Functions)

d) Hive Execution Engine

Ans - c) UDF (User-Defined Functions)

Hive uses a query language called:

a) Hadoop Query Language (HQL)

b) Hive Query Language (HiveQL)

c) SQL++

d) HiveSQL

Ans - b) Hive Query Language (HiveQL)

The schema of a Hive table is defined using:

a) Hive Query Language (HiveQL)

b) JSON files

c) XML files

d) Avro schemas

Ans - a) Hive Query Language (HiveQL)

Which of the following components in Hive is responsible for metadata management?

a) DataNode

b) NameNode

c) Hive Metastore

d) ResourceManager

Ans - c) Hive Metastore

Hive tables can be categorized into which two types based on how they store data?

a) Managed and External tables

b) Internal and External tables

c) Structured and Unstructured tables

d) Transactional and Non-transactional tables

Ans - b) Internal and External tables

Which type of table in Hive allows data to be stored in a directory outside the control of Hive?

a) Managed table

b) External table

c) Partitioned table

d) Bucketed table

Ans - b) External table

To improve query performance, Hive supports partitioning. What is the primary benefit of using partitioning in Hive?

a) It allows data to be stored in multiple formats.

b) It improves data security.

c) It enables data replication.

d) It reduces the amount of data scanned during query execution.

Ans - d) It reduces the amount of data scanned during query execution.

Which of the following is NOT a supported data storage format in Hive?

a) ORC (Optimized Row Columnar)

b) Parquet

c) Avro

d) CSV (Comma-Separated Values)

Ans - d) CSV (Comma-Separated Values)

What is the role of the HiveServer2 service in Hive?

a) It manages the Hive Metastore.

b) It serves as the query execution engine for Hive.

c) It provides an SQL interface to interact with Hive.

d) It handles data loading and exporting in Hive.

Ans - c) It provides an SQL interface to interact with Hive.

Which of the following commands is used to load data into a Hive table from an HDFS file?

a) INSERT INTO TABLE

b) UPDATE TABLE

c) LOAD DATA INFILE

d) LOAD DATA INPATH

Ans - d) LOAD DATA INPATH

In Hive, what does the term "partitioning" refer to?

a) Dividing a table into multiple smaller tables

b) Creating separate Hive instances for different users

c) Splitting data into smaller chunks for faster processing

d) Organizing data into subdirectories based on certain columns

Ans - d) Organizing data into subdirectories based on certain columns

What is the default storage format in Hive?

a) Parquet

b) ORC

c) Avro

d) TextFile

Ans - d) TextFile

In Hive, what is the role of the Metastore?

a) Storing the actual data files

b) Executing Hive queries

c) Managing metadata about tables and partitions

d) Performing data transformations using MapReduce

Ans - c) Managing metadata about tables and partitions

What language is used to write queries in Apache Hive?

a) Python

b) SQL

c) Java

d) HiveQL

Ans - d) HiveQL

In Hive, data is organized into which logical structure?

a) Tables

b) Databases

c) Partitions

d) Buckets

Ans - a) Tables

Which of the following is NOT a type of Hive table?

a) External table

b) Managed table

c) Temporary table

d) Transactional table

Ans - d) Transactional table

Hive supports two types of table formats for storage: ORC and _____?

a) HBase

b) Avro

c) Parquet

d) JSON

Ans - c) Parquet

What is the primary purpose of the Hive SerDe (Serializer/Deserializer)?

a) To manage table schemas

b) To perform data sorting in Hive

c) To serialize and deserialize data during data processing

d) To optimize Hive query execution plans

Ans - c) To serialize and deserialize data during data processing