MAPREDUCE

1. What is MapReduce?
- MapReduce is a programming model and software framework for processing large datasets in a parallel and distributed manner.
2. Who invented MapReduce?
- MapReduce was invented by Jeff Dean and Sanjay Ghemawat at Google.
3. What are the main components of MapReduce?
- The main components of MapReduce are the Map function, the Reduce function, and the Job Tracker.
4. What is the purpose of the Map function in MapReduce?
- The purpose of the Map function is to take an input dataset and convert it into a set of key-value pairs.
5. What is the purpose of the Reduce function in MapReduce?
- The purpose of the Reduce function is to take the output from the Map function and perform some kind of aggregation or computation on the data.
6. What is a key-value pair in MapReduce?
- A key-value pair is a data structure used to represent data in MapReduce. The key is a unique identifier for the data, and the value is the actual data itself.
7. What is the input format for MapReduce jobs?
- The input format for MapReduce jobs is typically a set of text files.
8. What is the output format for MapReduce jobs?
- The output format for MapReduce jobs is typically a set of text files.
9. What is the role of the Job Tracker in MapReduce?
- The Job Tracker is responsible for coordinating the execution of MapReduce jobs across a cluster of machines.
10. What is the role of the Task Tracker in MapReduce?
- The Task Tracker is responsible for executing individual tasks within a MapReduce job.
11. What is a task in MapReduce?
- A task is a unit of work that can be executed in parallel by a Task Tracker.
12. What is a map task in MapReduce?
- A map task is a task that processes a portion of the input data and produces a set of intermediate key-value pairs.
13. What is a reduce task in MapReduce?
- A reduce task is a task that processes a set of intermediate key-value pairs and produces a final set of output key-value pairs.
14. What is the Hadoop Distributed File System (HDFS)?
- HDFS is a distributed file system used by Hadoop and other distributed computing systems.
15. How does HDFS store data?

- HDFS stores data by breaking it into blocks and distributing those blocks across a cluster of machines.

16. How does MapReduce work with HDFS?
- MapReduce reads data from HDFS and writes output back to HDFS.

17. What is the role of a combiner function in MapReduce?
- A combiner function is an optional function that can be used to aggregate intermediate key-value pairs before they are sent to the reducers.

18. What is the purpose of shuffling in MapReduce?
- Shuffling is the process of sorting and transferring intermediate key-value pairs from the mappers to the reducers.

19. What is the default partitioner in MapReduce?
- The default partitioner in MapReduce is a hash partitioner.

20. What is a custom partitioner in MapReduce?
- A custom partitioner is a user-defined function that determines which reducer a given key-value pair will be sent to.

21. What is the purpose of speculative execution in MapReduce?
- Speculative execution is the process of launching duplicate tasks on multiple machines in order to improve job performance.

22. What is the purpose of input splits in MapReduce?
- Input splits are a way of dividing the input data into manageable chunks that can be processed in parallel.

23. What is a side effect in MapReduce?
- A side effect is any change to state outside of the

24. What is a distributed cache in MapReduce?
- A distributed cache is a mechanism for sharing files and data across a MapReduce cluster.

25. What is the purpose of the MapReduce programming model?
- The purpose of the MapReduce programming model is to provide a simple and scalable way to process large datasets.

26. What are some limitations of the MapReduce programming model?
- Some limitations of the MapReduce programming model include the need to write custom code for each job, the difficulty of expressing complex computations, and the lack of support for interactive queries.

27. What is the role of a Combiner in MapReduce?
- A Combiner is an optional function that can be used to reduce the amount of data that is transferred between the Map and Reduce phases.

28. What is the difference between a Mapper and a Reducer in MapReduce?
- A Mapper is responsible for processing the input data and emitting intermediate key-value pairs. A Reducer is responsible for processing the intermediate key-value pairs and producing the final output.

29. What is a partition in MapReduce?
- A partition is a unit of data that is assigned to a specific Reducer.

30. What is the purpose of the MapReduce shuffle phase?

- The purpose of the shuffle phase is to sort and transfer the intermediate key-value pairs from the Mappers to the Reducers.

31. What is the role of the MapReduce JobTracker?

- The JobTracker is responsible for managing the execution of MapReduce jobs across a cluster of machines.

32. What is the role of the MapReduce TaskTracker?

- The TaskTracker is responsible for executing individual tasks within a MapReduce job.

33. What is the difference between a MapReduce job and a Hadoop job?

- A MapReduce job is a specific type of Hadoop job that uses the MapReduce programming model.

34. What is the purpose of the Hadoop streaming API?

- The Hadoop streaming API allows users to write MapReduce jobs in any programming language that can read from stdin and write to stdout.

35. What is the purpose of the Hadoop pipes API?

- The Hadoop pipes API allows users to write MapReduce jobs in C++.

36. What is the purpose of the Hadoop MapReduce tutorial?

- The Hadoop MapReduce tutorial provides an introduction to the MapReduce programming model and the Hadoop ecosystem.

37. What is the purpose of the Hadoop MapReduce example programs?

- The Hadoop MapReduce example programs provide working examples of MapReduce jobs that demonstrate various aspects of the MapReduce programming model.

38. What is the purpose of the Hadoop streaming tutorial?

- The Hadoop streaming tutorial provides an introduction to writing MapReduce jobs using the Hadoop streaming API.

39. What is the purpose of the Hadoop pipes tutorial?

- The Hadoop pipes tutorial provides an introduction to writing MapReduce jobs using the Hadoop pipes API.

40. What is a YARN in Hadoop?

- YARN (Yet Another Resource Negotiator) is a component of Hadoop that is responsible for managing resources on a cluster and scheduling applications.

41. What is a cluster in Hadoop?

- A cluster is a group of machines that work together to process data using Hadoop.

42. What is the Hadoop ecosystem?

- The Hadoop ecosystem is a collection of open-source software projects that work together to provide a complete solution for processing large datasets.

43. What is the purpose of Apache Pig in Hadoop?

- Apache Pig is a platform for analyzing large datasets using a high-level language called Pig Latin.

44. What is the purpose of Apache Hive in Hadoop?

- Apache Hive is a data warehousing system that provides a SQL-like interface for querying data stored in Hadoop.

45. What is the purpose of Apache HBase in Hadoop?

- Apache HBase is a distributed, scalable, and fault-tolerant NoSQL database that is designed to store and manage large amounts of structured and semi-structured data.

46. What is the purpose of Apache Spark in Hadoop?

- Apache Spark is a fast and general-purpose distributed computing system that provides an interface for programming entire clusters with implicit data parallelism and fault tolerance.

47. What is the purpose of Apache Storm in Hadoop?

- Apache Storm is a distributed real-time processing system that is designed to process large volumes of streaming data.

48. What is the purpose of Apache Kafka in Hadoop?

- Apache Kafka is a distributed streaming platform that is designed to handle high volumes of real-time data feeds.

49. What is the purpose of Apache Flume in Hadoop?

- Apache Flume is a distributed, reliable, and available system for efficiently collecting, aggregating, and moving large amounts of log data from many different sources to a centralized data store.

50. What is the purpose of Apache Sqoop in Hadoop?

- Apache Sqoop is a tool designed to transfer data between Hadoop and structured data stores such as relational databases.

51. What is the purpose of Apache Zeppelin in Hadoop?

- Apache Zeppelin is a web-based notebook that provides an interactive environment for data exploration, visualization, and collaboration.

52. What is the purpose of Hadoop Distributed File System (HDFS)?

- Hadoop Distributed File System (HDFS) is a distributed file system designed to store large datasets across multiple machines.

53. What is the purpose of NameNode in HDFS?

- NameNode is the master node in HDFS that manages the file system namespace and controls access to files and directories.

54. What is the purpose of DataNode in HDFS?

- DataNode is the worker node in HDFS that stores and retrieves data from the file system.

55. What is the purpose of Secondary NameNode in HDFS?

- Secondary NameNode is a helper node in HDFS that periodically checkpoints the namespace and transaction logs of the NameNode to ensure faster recovery in case of failure.

56. What is the default replication factor in HDFS?

- The default replication factor in HDFS is 3.

57. What is the purpose of Block in HDFS?

- A Block in HDFS is a unit of data storage that is stored on a single DataNode.

58. What is the maximum file size that can be stored in HDFS?

- The maximum file size that can be stored in HDFS is determined by the block size and the number of blocks.

59. What is the difference between HDFS and traditional file systems?

- HDFS is designed to store and manage large datasets across multiple machines, while traditional file systems are designed to manage data on a single machine.

60. What is the purpose of Hadoop Common?

- Hadoop Common provides libraries and utilities that are used by other Hadoop components.

61. What is the purpose of JobTracker in Hadoop MapReduce?

- JobTracker is the master node in Hadoop MapReduce that manages and schedules jobs submitted to the cluster.

62. What is the purpose of TaskTracker in Hadoop MapReduce?

- TaskTracker is the worker node in Hadoop MapReduce that runs map and reduce tasks assigned by the JobTracker.

63. What is the difference between a map task and a reduce task in Hadoop MapReduce?

- A map task reads input data and generates intermediate key-value pairs, while a reduce task combines intermediate key-value pairs with the same key and generates output data.

64. What is the purpose of a combiner function in Hadoop MapReduce?

- A combiner function is a mini-reduce function that runs on the output of the map tasks before the data is sent to the reduce tasks. Its purpose is to reduce the amount of data transferred over the network and improve performance.

65. What is the purpose of a partitioner function in Hadoop MapReduce?

- A partitioner function determines which reducer gets which intermediate key-value pairs based on the key.

66. What is the purpose of a shuffle phase in Hadoop MapReduce?

- The shuffle phase in Hadoop MapReduce is the process of transferring intermediate key-value pairs from the map tasks to the reduce tasks.

67. What is the purpose of a sort phase in Hadoop MapReduce?

- The sort phase in Hadoop MapReduce is the process of sorting intermediate key-value pairs by key before they are passed to the reduce tasks.

68. What is the difference between a local mode and a cluster mode in Hadoop MapReduce?

- In local mode, Hadoop MapReduce runs on a single machine, while in cluster mode, it runs on a distributed cluster of machines.

69. What is the purpose of speculative execution in Hadoop MapReduce?

- Speculative execution in Hadoop MapReduce is the process of running multiple instances of the same task in parallel on different nodes to improve performance and reduce job completion time.

70. What is the purpose of Hadoop YARN?

- Hadoop YARN (Yet Another Resource Negotiator) is a resource management system that enables multiple data processing engines such as Hadoop MapReduce, Apache Spark, and Apache Tez to run on the same cluster.

71. What is the purpose of a container in Hadoop YARN?

- A container in Hadoop YARN is a resource allocation for running a specific task, such as a map or reduce task.

72. What is the difference between a node and a resource manager in Hadoop YARN?

- A node is a machine in a Hadoop YARN cluster that runs containers, while a resource manager is responsible for managing resources and scheduling containers on nodes.

73. What is the purpose of a scheduler in Hadoop YARN?

- A scheduler in Hadoop YARN is responsible for allocating resources to applications based on their needs and priorities.

74. What is the difference between capacity scheduler and fair scheduler in Hadoop YARN?

- Capacity scheduler allocates resources based on predefined capacities and queues, while fair scheduler dynamically allocates resources based on current demand and usage.

75. What is the purpose of a timeline server in Hadoop YARN?

- A timeline server in Hadoop YARN is responsible for storing and serving application-level metadata, such as job histories and application-specific metrics.

76. What is the purpose of Hadoop Oozie?

- Hadoop Oozie is a workflow scheduler system that is used to manage and run complex Hadoop jobs.

77. What is the purpose of Hadoop Hive?

- Hadoop Hive is a data warehousing and SQL-like query language that provides an interface to data stored in Hadoop.

78. What is the purpose of Hadoop Pig?

- Hadoop Pig is a high-level scripting language that is used to analyze and transform large datasets in Hadoop.

79. What is the purpose of Hadoop Mahout?

- Hadoop Mahout is a machine learning library that provides scalable algorithms for clustering, classification, and collaborative filtering.

80. What is the purpose of Hadoop Avro?

- Hadoop Avro is a data serialization system that provides a compact and efficient way of encoding data in Hadoop.

81. What is the purpose of Hadoop Flume?

- Hadoop Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data from various sources to a centralized data store.

82. What is the purpose of Hadoop Sqoop?

- Hadoop Sqoop is a tool that is used to transfer data between Hadoop and structured data stores such as relational databases.

83. What is the purpose of Hadoop HBase?

- Hadoop HBase is a distributed, non-relational database that provides real-time read/write access to large datasets in Hadoop.

84. What is the purpose of Hadoop ZooKeeper?

- Hadoop ZooKeeper is a distributed coordination service that is used to manage configuration information, provide distributed synchronization, and provide group services.

85. What is the purpose of Hadoop Accumulo?

- Hadoop Accumulo is a distributed, key/value store that is built on top of Hadoop and provides fine-grained access control and cell-level security.

86. What is the purpose of Hadoop Spark?

- Hadoop Spark is a fast and general-purpose data processing engine that is used for large-scale data processing, machine learning, and real-time stream processing.

87. What is the difference between Hadoop MapReduce and Hadoop Spark?

- Hadoop MapReduce is a batch processing system that is designed for processing large amounts of data, while Hadoop Spark is a data processing engine that is designed for real-time stream processing, machine learning, and interactive data analytics.

88. What is the purpose of Hadoop Flink?

- Hadoop Flink is a distributed data processing engine that is designed for high-throughput, low-latency, and fault-tolerant stream processing, batch processing, and graph processing.

89. What is the purpose of Hadoop Kafka?

- Hadoop Kafka is a distributed messaging system that is used for building real-time data pipelines and streaming applications.

90. What is the purpose of Hadoop NiFi?

- Hadoop NiFi is a data integration and distribution system that provides a web-based interface for designing, managing, and monitoring data flows in real-time.

91. What is the purpose of Hadoop Tez?

- Hadoop Tez is a distributed data processing framework that provides a generalized execution engine for building high-performance batch and interactive data processing applications.

92. What is the purpose of Hadoop Knox?

- Hadoop Knox is a security gateway system that provides secure access to Hadoop cluster resources using perimeter security controls.

93. What is the purpose of Hadoop Ranger?

- Hadoop Ranger is a security management framework that provides centralized authorization, auditing, and data protection for Hadoop clusters.

94. What is the purpose of Hadoop Atlas?

- Hadoop Atlas is a metadata management and governance system that provides a central repository for managing metadata related to data assets and their relationships.

95. What is the purpose of Hadoop YARN Timeline Service v2?

- Hadoop YARN Timeline Service v2 is a new version of the timeline server that provides improved scalability and performance for storing and serving application-level metadata in Hadoop YARN.

96. What is the purpose of Hadoop Distributed File System (HDFS)?

- Hadoop Distributed File System (HDFS) is the primary storage system for Hadoop and provides a distributed and scalable file system for storing large datasets across multiple machines.

97. What is the difference between Hadoop HDFS and a traditional file system?

- Hadoop HDFS is designed for handling large, unstructured data sets across many machines, while traditional file systems are designed for handling smaller, structured data sets on a single machine.

98. What is the purpose of Hadoop NameNode?

- Hadoop NameNode is the central node in the Hadoop HDFS architecture and is responsible for managing the file system namespace, metadata, and access control.

99. What is the purpose of Hadoop DataNode?

- Hadoop DataNode is a node in the Hadoop HDFS architecture that is responsible for storing and retrieving data from the file system.

100. What is the purpose of Hadoop Secondary NameNode?

- Hadoop Secondary NameNode is a node in the Hadoop HDFS architecture that is responsible for periodically merging the namespace and edits files to create a new checkpoint of the file system metadata. It is not a backup or a replacement for the NameNode.

MCQ-: MapReduce

What is MapReduce?
A. A database management system
B. A distributed computing framework
C. A file system

D. A programming language
Which of the following is not a key feature of MapReduce?
A. Scalability
B. Fault-tolerance
C. High-availability

D. Low-latency
In MapReduce, what is the role of the mapper?
A. To reduce the data
B. To sort the data
C. To transform the data

D. To shuffle the data
In MapReduce, what is the role of the reducer?
A. To reduce the data
B. To sort the data
C. To transform the data

D. To shuffle the data
Which of the following is not a component of MapReduce?
A. JobTracker
B. TaskTracker
C. NameNode

D. DataNode
In MapReduce, which component is responsible for assigning tasks to individual nodes?
A. JobTracker
B. TaskTracker
C. NameNode

D. DataNode
In MapReduce, which component is responsible for managing the overall job execution?

A. JobTracker
B. TaskTracker
C. NameNode

D. DataNode
In MapReduce, which component is responsible for storing the input and output data?
A. JobTracker
B. TaskTracker
C. NameNode

D. DataNode
Which of the following is a disadvantage of MapReduce?
A. Low latency
B. Complex programming model
C. Limited scalability

D. No fault-tolerance
Which of the following is a key advantage of MapReduce?
A. High performance on small data sets
B. Ability to process real-time data
C. Ability to scale to large data sets

D. Simple programming model
Which of the following is not a common use case for MapReduce?
A. Log analysis
B. Sentiment analysis
C. Image processing

D. Fraud detection
Which programming language is commonly used to write MapReduce jobs?
A. Java
B. Python
C. Ruby

D. JavaScript
In MapReduce, what is the input format?
A. The format in which data is stored in HDFS
B. The format in which data is passed to the mapper
C. The format in which data is passed to the reducer

D. The format in which data is stored in a database
In MapReduce, what is the output format?
A. The format in which data is stored in HDFS
B. The format in which data is passed to the mapper

C. The format in which data is passed to the reducer

D. The format in which data is stored in a database
In MapReduce, what is the partitioner?
A. The component that divides the input data into chunks
B. The component that assigns tasks to individual nodes
C. The component that sorts the intermediate key-value pairs

D. The component that determines which reducer receives which intermediate data
In MapReduce, what is the combiner?
A. The component that divides the input data into chunks
B. The component that assigns tasks to individual nodes
C. The component that sorts the intermediate key-value pairs

D. The component that performs local aggregation on the intermediate data
In MapReduce, what is the shuffle phase?
A. The phase in which data is divided into chunks
B. The phase in which tasks are assigned to individual nodes
C. The phase in which intermediate data is sorted and grouped

D. The phase

Which of the following is not a benefit of using MapReduce?
A. High scalability
B. Fault tolerance
C. Low programming complexity
D. Low cost
What does Hadoop provide to implement MapReduce?
A. A programming language
B. A distributed file system
C. A database management system
D. A data visualization tool
Which of the following is not a component of Hadoop?
A. YARN
B. HDFS
C. Pig
D. ZooKeeper
In Hadoop, what is YARN?
A. A resource manager for distributed applications
B. A distributed file system
C. A programming language for MapReduce
D. A data processing framework

Which of the following is a key advantage of Hadoop MapReduce over traditional data processing systems?
A. Ability to handle real-time data processing
B. Lower cost
C. Lower programming complexity
D. Lower storage requirements
In Hadoop, what is the role of the NameNode?
A. To manage the file system metadata
B. To manage the data storage on individual nodes
C. To manage the job execution
D. To manage the task execution
In Hadoop, what is the role of the DataNode?
A. To manage the file system metadata
B. To manage the data storage on individual nodes
C. To manage the job execution
D. To manage the task execution
Which of the following is a commonly used tool for developing and testing Hadoop MapReduce jobs?
A. Eclipse
B. Visual Studio
C. Sublime Text
D. Notepad++
Which of the following is a commonly used tool for monitoring and managing Hadoop clusters?
A. Ambari
B. Jenkins
C. GitLab
D. JIRA
In Hadoop, what is the role of the ResourceManager?
A. To manage the file system metadata
B. To manage the data storage on individual nodes
C. To manage the job execution
D. To manage the task execution
In Hadoop, what is the role of the NodeManager?
A. To manage the file system metadata
B. To manage the data storage on individual nodes
C. To manage the job execution
D. To manage the task execution
Which of the following is not a commonly used data serialization format in Hadoop?
A. JSON
B. XML
C. Protocol Buffers
D. YAML
In Hadoop, which of the following is not a commonly used data processing framework?
A. Pig

B. Hive
C. Spark
D. Flask
In Hadoop, what is the role of the job tracker?
A. To manage the file system metadata
B. To manage the data storage on individual nodes
C. To manage the job execution
D. To manage the task execution
In Hadoop, what is the role of the task tracker?
A. To manage the file system metadata
B. To manage the data storage on individual nodes
C. To manage the job execution
D. To manage the task execution
Which of the following is not a commonly used method for input data partitioning in Hadoop?
A. Hash partitioning
B. Range partitioning
C. Round-robin partitioning
D. Binary partitioning
In Hadoop, what is the maximum number of reducers that can be used in a MapReduce job?
A. 100
B. 1,000
C. 10,000
D. There is no maximum limit
Which of the following is not a commonly used method for data aggregation

In Hadoop MapReduce, what is the output format of a map task?
A. Key-value pairs
B. Text file
C. XML file
D. Binary file
In Hadoop MapReduce, what is the output format of a reduce task?
A. Key-value pairs
B. Text file
C. XML file
D. Binary file
Which of the following is not a commonly used join operation in Hadoop MapReduce?
A. Inner join
B. Outer join
C. Left join
D. Full join
In Hadoop MapReduce, which of the following is not a commonly used type of input format?
A. TextInputFormat
B. KeyValueInputFormat
C. SequenceFileInputFormat

D. ImageInputFormat
In Hadoop MapReduce, which of the following is not a commonly used type of output format?
A. TextOutputFormat
B. KeyValueOutputFormat
C. SequenceFileOutputFormat
D. ImageOutputFormat
Which of the following is not a commonly used tool for managing Hadoop clusters?
A. Ambari
B. Cloudera Manager
C. Hortonworks Data Platform
D. Apache Spark
In Hadoop, what is the role of the Secondary NameNode?
A. To provide backup for the NameNode
B. To manage the data storage on individual nodes
C. To manage the job execution
D. To manage the task execution
In Hadoop MapReduce, what is the purpose of the combiner function?
A. To aggregate the intermediate values generated by the map task before sending them to the reduce task
B. To aggregate the output of multiple reduce tasks
C. To sort the intermediate values generated by the map task before sending them to the reduce task
D. To filter the input data before passing it to the map task
Which of the following is not a commonly used data source for Hadoop MapReduce jobs?
A. HDFS
B. Local file system
C. Amazon S3
D. Microsoft SQL Server
In Hadoop MapReduce, which of the following is not a commonly used type of data join?
A. Map-side join
B. Reduce-side join
C. Distributed cache join
D. Inner join
In Hadoop, what is the purpose of the DistributedCache?
A. To provide a shared file system for Hadoop clusters
B. To cache the output of map tasks for future use
C. To cache frequently accessed files in memory for faster processing
D. To cache small amounts of data on the nodes in the cluster
Which of the following is not a commonly used tool for data analysis in Hadoop?
A. Apache Pig
B. Apache Hive
C. Apache Spark
D. Apache Tomcat
In Hadoop MapReduce, which of the following is not a commonly used type of partitioner?
A. HashPartitioner

B. RangePartitioner
C. KeyFieldBasedPartitioner
D. KeyValuePartitioner
Which of the following is not a commonly used tool for data visualization in Hadoop?
A. Tableau
B. Apache Zeppelin
C. QlikView
D. Apache Kafka
In Hadoop MapReduce, what is the purpose of the JobConf object?
A. To configure the job properties
B. To submit the job to the cluster
C. To define the map and reduce tasks
D. To monitor the job execution status.

In Hadoop MapReduce, what is the purpose of the TaskTracker?
A. To execute map and reduce tasks on individual nodes
B. To manage the data storage on individual nodes
C. To manage the job execution
D. To provide backup for the NameNode
In Hadoop MapReduce, which of the following is not a commonly used type of combiner function?
A. SumCombiner
B. AverageCombiner
C. MaxCombiner
D. MinCombiner
In Hadoop MapReduce, which of the following is not a commonly used type of output collector?
A. KeyValueOutputCollector
B. RecordWriter
C. TextOutputCollector
D. SequenceFileOutputCollector
In Hadoop MapReduce, what is the purpose of the InputSplit?
A. To divide the input data into manageable chunks for processing
B. To provide backup for the NameNode
C. To manage the job execution
D. To manage the task execution
In Hadoop MapReduce, which of the following is not a commonly used type of reducer function?
A. CountReducer
B. AverageReducer
C. MaxReducer
D. MinReducer
Which of the following is not a commonly used tool for scheduling jobs in Hadoop?
A. Oozie
B. Azkaban
C. Airflow

D. Apache Kafka
In Hadoop MapReduce, what is the purpose of the RecordReader?
A. To read input data from a data source and produce key-value pairs for the map task
B. To aggregate the output of multiple reduce tasks
C. To sort the intermediate values generated by the map task before sending them to the reduce task
D. To filter the input data before passing it to the map task
In Hadoop, what is the role of the ResourceManager?
A. To manage the task execution
B. To manage the data storage on individual nodes
C. To provide backup for the NameNode
D. To manage the resource allocation for the cluster
In Hadoop MapReduce, which of the following is not a commonly used type of input split?
A. TextInputSplit
B. SequenceFileInputSplit
C. KeyValueInputSplit
D. ImageInputSplit
In Hadoop MapReduce, what is the purpose of the partitioner?
A. To partition the output of the map task for processing by the reduce task
B. To aggregate the intermediate values generated by the map task before sending them to the reduce task
C. To sort the intermediate values generated by the map task before sending them to the reduce task
D. To filter the input data before passing it to the map task

In Hadoop MapReduce, which of the following is not a commonly used type of mapper function?
A. IdentityMapper
B. TokenizerMapper
C. CombinerMapper
D. WordCountMapper
In Hadoop MapReduce, what is the role of the JobTracker?
A. To manage the job execution
B. To manage the data storage on individual nodes
C. To provide backup for the NameNode
D. To manage the task execution
Which of the following is not a commonly used framework for implementing MapReduce jobs in Java?
A. Apache Hadoop
B. Apache Spark
C. Apache Flink
D. Apache Cassandra
In Hadoop MapReduce, which of the following is not a commonly used type of input format?
A. TextInputFormat
B. SequenceFileInputFormat
C. KeyValueInputFormat

D. ImageInputFormat
In Hadoop MapReduce, which of the following is not a commonly used type of output format?
A. TextOutputFormat
B. SequenceFileOutputFormat
C. KeyValueOutputFormat
D. ImageOutputFormat
In Hadoop MapReduce, what is the purpose of the JobConf object?
A. To configure the MapReduce job
B. To manage the resource allocation for the cluster
C. To manage the task execution
D. To manage the data storage on individual nodes
Which of the following is not a commonly used method for optimizing MapReduce jobs?
A. Using combiner functions
B. Using partitioners
C. Using a large number of small input files
D. Using compression
In Hadoop MapReduce, what is the purpose of the MapReduce API?
A. To provide a programming interface for MapReduce jobs
B. To manage the resource allocation for the cluster
C. To manage the data storage on individual nodes
D. To manage the task execution
In Hadoop MapReduce, which of the following is not a commonly used type of data format for input data?
A. Text
B. CSV
C. JSON
D. MP4
In Hadoop MapReduce, which of the following is not a commonly used type of data format for output data?
A. Text
B. CSV
C. JSON
D. MP4

In Hadoop MapReduce, what is the purpose of the Reducer function?
A. To perform the map phase
B. To aggregate and process the intermediate key-value pairs generated by the mapper function
C. To distribute the data across the nodes in the cluster
D. To manage the resource allocation for the cluster
In Hadoop MapReduce, which of the following is not a commonly used method for handling missing data in input files?
A. Ignoring the missing data
B. Replacing the missing data with an average or median value
C. Removing the entire row containing the missing data

D. Replacing the missing data with a random value
In Hadoop MapReduce, which of the following is not a commonly used method for handling skew in the input data?
A. Using a partitioner to distribute the input data evenly across the nodes in the cluster
B. Using a combiner function to aggregate the intermediate key-value pairs generated by the mapper function
C. Using a custom partitioner to partition the input data based on the skew
D. Using a custom reducer function to handle the skew in the input data
In Hadoop MapReduce, which of the following is not a commonly used technique for improving the performance of the MapReduce job?
A. Increasing the number of mapper and reducer tasks
B. Using a larger cluster
C. Using more powerful hardware
D. Using a more complex algorithm for the mapper and reducer functions
In Hadoop MapReduce, what is the purpose of the OutputCollector object?
A. To collect and aggregate the intermediate key-value pairs generated by the mapper function
B. To distribute the data across the nodes in the cluster
C. To manage the resource allocation for the cluster
D. To collect and write the output key-value pairs to the output file
In Hadoop MapReduce, what is the purpose of the DistributedCache feature?
A. To cache intermediate data generated by the mapper and reducer functions
B. To distribute files and archives needed by the mapper and reducer functions
C. To manage the resource allocation for the cluster
D. To distribute the data across the nodes in the cluster
In Hadoop MapReduce, which of the following is not a commonly used technique for reducing the amount of data shuffled across the network?
A. Using a combiner function
B. Using a partitioner function
C. Using a custom serializer
D. Using a larger number of mapper tasks
In Hadoop MapReduce, what is the purpose of the InputFormat class?
A. To provide a way of reading input data from different sources
B. To configure the MapReduce job
C. To manage the resource allocation for the cluster
D. To manage the data storage on individual nodes
In Hadoop MapReduce, which of the following is not a commonly used method for handling outliers in input data?
A. Removing the outliers from the input data
B. Replacing the outliers with an average or median value
C. Keeping the outliers and treating them as separate data points
D. Replacing the outliers with a random value
In Hadoop MapReduce, which of the following is not a commonly used technique for improving the fault tolerance of the MapReduce job?
A. Using speculative execution

B. Using backup nodes for the NameNode
C. Using redundant storage for the HDFS
D. Using checkpointing and rollback

In Hadoop MapReduce, which of the following is not a commonly used method for handling text input data?
A. Using TextInputFormat
B. Using SequenceFileInputFormat
C. Using AvroInputFormat
D. Using XMLInputFormat
In Hadoop MapReduce, what is the purpose of the JobTracker?
A. To manage the HDFS
B. To manage the MapReduce jobs
C. To manage the cluster resources
D. To manage the data storage on individual nodes
In Hadoop MapReduce, what is the purpose of the TaskTracker?
A. To manage the HDFS
B. To manage the MapReduce jobs
C. To manage the cluster resources
D. To manage the individual mapper and reducer tasks
In Hadoop MapReduce, what is the purpose of the Combiner function?
A. To aggregate and process the intermediate key-value pairs generated by the mapper function
B. To distribute the data across the nodes in the cluster
C. To manage the resource allocation for the cluster
D. To manage the individual mapper and reducer tasks
In Hadoop MapReduce, which of the following is not a commonly used method for handling binary input data?
A. Using SequenceFileInputFormat
B. Using AvroInputFormat
C. Using TextInputFormat
D. Using KeyValueInputFormat
In Hadoop MapReduce, what is the purpose of the RecordReader class?
A. To read the input data and split it into key-value pairs
B. To aggregate and process the intermediate key-value pairs generated by the mapper function
C. To distribute the data across the nodes in the cluster
D. To manage the resource allocation for the cluster
In Hadoop MapReduce, what is the purpose of the SequenceFileOutputFormat class?
A. To provide a way of writing output data in a sequence file format
B. To provide a way of writing output data in a text file format
C. To manage the resource allocation for the cluster
D. To distribute the data across the nodes in the cluster
In Hadoop MapReduce, what is the purpose of the Partitioner function?
A. To distribute the data across the nodes in the cluster
B. To aggregate and process the intermediate key-value pairs generated by the mapper function

C. To manage the resource allocation for the cluster
D. To partition the output data based on the key to be written to the output file
In Hadoop MapReduce, what is the purpose of the Mapper function?
A. To read the input data and split it into key-value pairs
B. To aggregate and process the intermediate key-value pairs generated by the reducer function
C. To distribute the data across the nodes in the cluster
D. To perform some transformation on the input data and emit intermediate key-value pairs
In Hadoop MapReduce, which of the following is not a commonly used method for handling categorical input data?
A. One-Hot Encoding
B. Binary Encoding
C. Label Encoding
D. Z-score normalization

In Hadoop MapReduce, what is the purpose of the Reducer function?
A. To read the input data and split it into key-value pairs
B. To aggregate and process the intermediate key-value pairs generated by the mapper function
C. To distribute the data across the nodes in the cluster
D. To perform some final aggregation on the intermediate key-value pairs and write the output data

In Hadoop MapReduce, which of the following is a commonly used data serialization framework?
A. XML
B. CSV
C. Avro
D. JSON
In Hadoop MapReduce, which of the following is not a commonly used method for handling output data?
A. Using TextOutputFormat
B. Using SequenceFileOutputFormat
C. Using AvroOutputFormat
D. Using BinaryOutputFormat
In Hadoop MapReduce, what is the purpose of the JobConf class?
A. To configure the input and output formats for a MapReduce job
B. To configure the resource allocation for the cluster
C. To manage the individual mapper and reducer tasks
D. To manage the HDFS
In Hadoop MapReduce, what is the purpose of the Counters feature?
A. To provide a way of measuring the progress of a MapReduce job
B. To provide a way of counting the number of nodes in the cluster
C. To provide a way of counting the number of mapper and reducer tasks
D. To provide a way of counting the number of input and output records
In Hadoop MapReduce, which of the following is not a commonly used method for handling output compression?

A. Using GzipCodec
B. Using LzoCodec
C. Using Bzip2Codec
D. Using TextCodec
In Hadoop MapReduce, what is the purpose of the DistributedCache feature?
A. To provide a way of sharing files among the nodes in the cluster
B. To provide a way of distributing the input data across the nodes in the cluster
C. To provide a way of distributing the output data across the nodes in the cluster
D. To provide a way of distributing the MapReduce job among the nodes in the cluster
In Hadoop MapReduce, which of the following is a commonly used method for handling large input data?
A. Using SequenceFileInputFormat
B. Using AvroInputFormat
C. Using TextInputFormat
D. Using CombineTextInputFormat
In Hadoop MapReduce, which of the following is not a commonly used method for handling output data compression?
A. Using GzipCodec
B. Using LzoCodec
C. Using Bzip2Codec
D. Using SnappyCodec
In Hadoop MapReduce, which of the following is not a commonly used method for handling unstructured input data?
A. Using TextInputFormat
B. Using SequenceFileInputFormat
C. Using AvroInputFormat
D. Using KeyValueInputFormat

…………………………………………………………………………….