CMT 428 Module 5: Advanced Hadoop Topics Quiz

Questions

1. What is the primary function of YARN in a Hadoop ecosystem? a) Data storage b) Data processing c) Resource management d) Security management
2. Which of the following is NOT a benefit of using YARN? a) Scalability b) Multi-tenancy c) Improved data compression d) Resource utilization
3. Spark's in-memory processing capability makes it significantly faster than traditional MapReduce for which type of computations? a) Single-pass computations b) Iterative computations c) Sequential computations d) Batch computations
4. What is the primary advantage of using Spark over MapReduce for data analysis? a) Reduced storage space b) Increased fault tolerance c) Faster processing speed d) Simplified data modeling
5. Which Spark component is responsible for scheduling and executing tasks across the cluster? a) Spark Driver b) Spark Executor c) Spark Master d) Spark Worker
6. What is the core concept behind Kafka's architecture? a) Distributed queues b) Key-value stores c) Relational databases d) Graph databases
7. Which of the following is NOT a key characteristic of Kafka? a) High throughput b) Fault tolerance c) In-memory processing d) Scalability
8. Kafka is typically used for which of the following use cases? a) Batch data processing b) Real-time data streaming c) Static data storage d) Data visualization
9. What is the role of a Kafka Producer? a) Consume messages from topics b) Store messages in topics c) Publish messages to topics d) Manage the Kafka cluster
10. What is the purpose of replicating data blocks in HDFS? a) Improve data security b) Increase storage capacity c) Enhance data locality d) Ensure fault tolerance
11. Which parameter in HDFS determines the size of individual data blocks? a) dfs.block.size b) dfs.replication c) yarn.nodemanager.resource.memory-mb d) mapreduce.map.memory.mb
12. What is the recommended approach for optimizing the number of mappers in a MapReduce job? a) Use as many mappers as possible b) Use a single mapper for the entire dataset c) Match the number of mappers to the number of input splits d) Minimize the number of mappers to reduce overhead
13. Which tool can be used to monitor and analyze the performance of a Hadoop cluster? a) Eclipse b) IntelliJ IDEA c) Ganglia d) Visual Studio Code
14. What is the purpose of data compression in Hadoop? a) Improve data security b) Reduce storage space c) Increase processing speed d) Simplify data modeling
15. Which of the following compression codecs is commonly used in Hadoop? a) JPEG b) MP3 c) Snappy d) ZIP

16. What is the function of the NodeManager in YARN? a) Manage resources for individual nodes b) Schedule jobs and allocate resources c) Monitor the health of the cluster d) Execute application code

17. What does the term "speculative execution" refer to in the context of MapReduce? a) Running multiple copies of a task to mitigate slow nodes b) Predicting the outcome of a task before it completes c) Optimizing task execution based on data locality d) Executing tasks in a random order to avoid biases

18. Which of the following is a valid file format for storing data in HDFS? a) CSV b) JPEG c) MP4 d) All of the above

19. What is the role of ZooKeeper in a Kafka cluster? a) Store message data b) Manage consumer offsets c) Maintain cluster metadata d) Authenticate client connections

20. Which API in Spark is used for processing streaming data? a) Spark SQL b) Spark Streaming c) Spark MLlib d) Spark Core

Answers

1. c
2. c
3. b
4. c
5. a
6. a
7. c
8. b
9. c
10. d
11. a
12. c
13. c
14. b
15. c
16. a
17. a
18. d
19. c
20. b