CMT 428 Final Quiz

Module 1: Introduction to Big Data and Hadoop

1. What is a primary challenge in Big Data?
    - a) Data accuracy
    - b) Data variety
    - c) Data analysis speed
    - d) Data storage costs
2. Which component is the core storage system of Hadoop?
    - a) MapReduce
    - b) HDFS
    - c) Spark
    - d) Sqoop
3. The Hadoop ecosystem primarily addresses:
    - a) Financial data processing
    - b) Real-time data analysis
    - c) Large-scale data storage and processing
    - d) Multimedia data management
4. Which Hadoop component is responsible for processing data?
    - a) HDFS
    - b) Oozie
    - c) MapReduce
    - d) Flume
5. Which of these is a characteristic of Big Data?
    - a) Volume
    - b) Simplicity
    - c) Accuracy
    - d) Flexibility
6. HDFS is optimized for:
    - a) High-speed networking
    - b) Real-time data transactions
    - c) Large files and sequential data access

- d) Multimedia content

7. In Hadoop, data redundancy is achieved through:
    - a) Compression algorithms
    - b) Replication in HDFS
    - c) Data encryption
    - d) File sharding
8. Which is a core function of the Hadoop Distributed File System (HDFS)?
    - a) Data streaming
    - b) Job scheduling
    - c) Fault tolerance
    - d) User authentication

Module 2: Hadoop Cluster Setup and Management

9. Which command initiates Hadoop installation?
    - a) hadoop init
    - b) start-hadoop
    - c) hadoop-install
    - d) start-all.sh
10. Hadoop cluster management primarily involves:
    - a) Data replication only
    - b) Monitoring and maintaining cluster health
    - c) Developing data models
    - d) Setting up networking layers
11. Which tool in Hadoop helps monitor cluster performance?
    - a) HBase
    - b) Hive
    - c) Oozie
    - d) Hadoop Metrics
12. The primary configuration files in Hadoop are:
    - a) core-site.xml, hdfs-site.xml, mapred-site.xml
    - b) config.xml, hadoop-site.xml, cluster-site.xml
    - c) hadoop-env.xml, system.xml, data.xml

- d) core.xml, spark-site.xml, cluster.xml

13. Which is NOT part of Hadoop's default security model?
    - a) Data encryption
    - b) User authentication
    - c) Role-based access control
    - d) Replication
14. Hadoop's cluster architecture is designed to:
    - a) Prioritize security over performance
    - b) Handle multiple small files efficiently
    - c) Work on low-cost commodity hardware
    - d) Support real-time processing natively
15. In a Hadoop cluster, the NameNode is responsible for:
    - a) Storing data blocks
    - b) Managing file system metadata
    - c) Running MapReduce tasks
    - d) Scheduling resources
16. The purpose of Hadoop's Secondary NameNode is:
    - a) Data redundancy
    - b) Backup of the NameNode metadata
    - c) Running MapReduce jobs
    - d) Managing data encryption

Module 3: MapReduce Programming

17. The purpose of the Map function in MapReduce is to:
    - a) Sort data
    - b) Filter data
    - c) Break down data into key-value pairs
    - d) Aggregate data
18. Which component handles the execution of a MapReduce job?
    - a) Task Tracker
    - b) Job Tracker
    - c) YARN Resource Manager

- d) DataNode

19. MapReduce is primarily designed for:
    - a) Real-time processing
    - b) Batch processing
    - c) Streaming data
    - d) Data visualization
20. In a MapReduce job, the Reduce phase is used to:
    - a) Filter data
    - b) Sort data
    - c) Aggregate and summarize results
    - d) Shuffle data
21. Which optimization strategy minimizes data movement in MapReduce?
    - a) Indexing
    - b) Sorting
    - c) Localizing data processing to nodes
    - d) Increasing data replication
22. MapReduce applications are typically written in:
    - a) C++
    - b) Python
    - c) Java
    - d) HTML
23. A combiner in MapReduce helps by:
    - a) Filtering unnecessary data
    - b) Minimizing data transfer between Map and Reduce phases
    - c) Sorting data
    - d) Aggregating final results
24. Which is an example of a MapReduce design pattern?
    - a) Fork-Join
    - b) Shuffle-Sort
    - c) Search-Sort
    - d) Mapper-Reducer

Module 4: Hadoop Ecosystem Tools

25. Which Hadoop tool is primarily used for data warehousing?
    - a) Pig
    - b) HBase
    - c) Hive
    - d) Oozie
26. Pig Latin is associated with:
    - a) Hive
    - b) Pig
    - c) HBase
    - d) Flume
27. HBase is best described as a:
    - a) NoSQL database
    - b) Real-time analytics tool
    - c) Data import tool
    - d) Data warehousing tool
28. Sqoop is used in Hadoop for:
    - a) Data processing
    - b) Data import/export
    - c) Data security
    - d) Data replication
29. Flume is primarily responsible for:
    - a) Running batch jobs
    - b) Real-time data streaming
    - c) Job scheduling
    - d) File replication
30. Oozie is a tool for:
    - a) Data analytics
    - b) Workflow management
    - c) NoSQL storage
    - d) Data security

31. Hive supports querying data using:
    - a) Pig Latin
    - b) SQL-like syntax
    - c) JSON
    - d) YAML
32. The HBase data model is based on:
    - a) Key-value pairs
    - b) Relational tables
    - c) Hierarchical structure
    - d) JSON format

Module 5: Advanced Hadoop Topics

33. YARN in Hadoop is primarily used for:
    - a) Data replication
    - b) Resource management
    - c) File organization
    - d) Security
34. Spark is preferred over MapReduce for:
    - a) Data warehousing
    - b) Real-time, in-memory processing
    - c) Batch processing only
    - d) File system organization
35. Kafka is used in Hadoop for:
    - a) Real-time data streaming
    - b) Data replication
    - c) Workflow management
    - d) Query processing
36. Tuning Hadoop performance is essential to:
    - a) Reduce storage space
    - b) Maximize job execution efficiency
    - c) Enhance user interface
    - d) Limit data access

37. Which of the following is a method of performance optimization in Hadoop?
    - a) Reducing data block size
    - b) Increasing replication factor
    - c) Using compression
    - d) Decreasing node count
38. Apache Spark processes data in:
    - a) Real-time
    - b) Sequential batches
    - c) Text format only
    - d) Network layers
39. What does YARN stand for?
    - a) Yet Another Resource Negotiator
    - b) Yielding Analysis Resource Node
    - c) Yearly Access Resource Network
    - d) Yet Another Resource Network
40. Which framework supports in-memory processing?
    - a) Oozie
    - b) Spark
    - c) Hive
    - d) HBase

Answers

1. b) Data variety
2. b) HDFS
3. c) Large-scale data storage and processing
4. c) MapReduce
5. a) Volume
6. c) Large files and sequential data access
7. b) Replication in HDFS
8. c) Fault tolerance
9. d) start-all.sh
10. b) Monitoring and maintaining cluster health
11. d) Hadoop Metrics
12. a) core-site.xml, hdfs-site.xml, mapred-site.xml
13. a) Data encryption
14. c) Work on low-cost commodity hardware
15. b) Managing file system metadata
16. b) Backup of the NameNode metadata
17. c) Break down data into key-value pairs
18. b) Job Tracker
19. b) Batch processing
20. c) Aggregate and summarize results
21. c) Localizing data processing to nodes
22. c) Java
23. b) Minimizing data transfer between Map and Reduce phases
24. d) Mapper-Reducer
25. c) Hive
26. b) Pig
27. a) NoSQL database
28. b) Data import/export
29. b) Real-time data streaming
30. b) Workflow management

31. b) SQL-like syntax
32. a) Key-value pairs
33. b) Resource management
34. b) Real-time, in-memory processing
35. a) Real-time data streaming
36. b) Maximize job execution efficiency
37. c) Using compression
38. a) Real-time
39. a) Yet Another Resource Negotiator
40. b) Spark