Reg. No															

B.Tech DEGREE EXAMINATION, MAY 2024

Seventh Semester

18CSE333J - BIG DATA TOOLS AND TECHNIQUES FOR BLOCKCHAIN

(For the candidates admitted during the academic year 2018-2019 to 2021-2022)

Note:

i. **Part - A** should be answered in OMR sheet within first 40 minutes and OMR sheet should be handed over to hall invigilator at the end of 40th minute.
ii. **Part - B** and **Part - C** should be answered in answer booklet.

Time: 3 Hours **Max. Marks: 100**

PART - A (20 × 1 = 20 Marks)

Answer **all** Questions

	Question	Marks	BL	CO
1.	Which Hadoop component is responsible for the distributed storage of data? (A) Hadoop MapReduce (B) Hadoop HDFS (C) Hadoop Pig (D) Hadoop Hive	1	2	1
2.	What does the term "Map" phase refer to in the MapReduce programming model? (A) for transformine or mapping input to output key value pairs (B) Data sorting and shuffling (C) Data storage in HDFS (D) Data retrieval from HDFS	1	2	1
3.	Which Hadoop ecosystem project provides a SQL-like interface for querying and analyzing data stored in Hadoop? (A) Sqoop (B) Pig (C) Flume (D) Spark	1	2	1
4.	In Hadoop, what is the purpose of the "Reducer" phase in the MapReduce programming model? (A) Data splitting and mapping (B) Data aggregation and summarization (C) Data storage in HDFS (D) Data retrieval from HDFS	1	2	1
5.	In HDFS, what is the maximum file size that can be stored by default? (A) 1 terabyte (B) 2 terabytes (C) 4 terabytes (D) 16 terabytes	1	4	2
6.	How does HDFS ensure fault tolerance for data storage? (A) By using checksums for data verification (B) By compressing data blocks for efficient storage (C) By replicating data blocks across multiple DataNodes (D) B y encrypting data at rest	1	4	2
7.	What is the purpose of the "Block Report" in HDFS? (A) To report the status of DataNodes to the NameNode (B) To report the status of the secondary NameNode to the ResourceManager (C) To report the status of the NameNode to DataNodes (D) To report the status of the ResourceManager to the NameNode	1	5	2
8.	How does HDFS handle data locality to optimize data processing? (A) By dynamically replicating data to achieve locality (B) By randomly distributing data across DataNodes (C) By using a centralized data processing approach (D) By using a single, large DataNode for all data storage	1	5	2

9. The default number of copies of each block in HDFS is ___________ 1 2 3

(A) 1 (B) 2
(C) 3 (D) 4

10. What is the output format of the Map phase in MapReduce? 1 2 3

(A) Key-value pairs (B) Comma-separated values
(C) Tab-separated values (D) JSON format

11. In a MapReduce job, what does the Shuffle and Sort phase primarily involve? 1 2 3

(A) Combining data from multiple Map tasks (B) Sorting and grouping data by key
(C) Reducing data size before storage (D) Distributing data across the cluster

12. __________________component in map reduce programming paradigm is used to minimize the number of key value pairs transferred between map and reduce phases. 1 2 3

(A) combiner (B) CRC32C
(C) Serializer (D) Shuffle

13. Which Hadoop ecosystem component is commonly used for storing and querying large-scale, semi-structured data? 1 4 4

(A) HDFS (B) Sqoop
(C) HBase (D) Flume

14. What is the primary role of Apache Kafka in the Hadoop ecosystem? 1 4 4

(A) Real-time data processing (B) Storing structured data
(C) Batch processing of large datasets (D) Data visualization

15. What is the purpose of Apache Ambari in the Hadoop ecosystem? 1 2 4

(A) To manage and monitor Hadoop clusters (B) To query and analyze data in real-time
(C) To build and execute machine learning models (D) To manage distributed file systems

16. Which Hadoop ecosystem component is designed for real-time, interactive querying and analysis of data? 1 4 4

(A) Pig (B) Impala
(C) YARN (D) Mahout

17. What is the primary goal of data analytics? 1 2 5

(A) To predict future events accurately (B) To summarize historical data
(C) To develop web applications (D) To develop collaborative filtering algorithms

18. In machine learning, what does "supervised learning" involve? 1 2 5

(A) Training a model with labeled data to make predictions (B) Learning from data without any guidance
(C) Collaborating with other machine learning models (D) Analyzing unstructured text data

19. Which of the following is an example of an unsupervised learning technique? 1 3 5

(A) Decision trees (B) k-Means clustering
(C) Linear regression (D) Random forests

20. What is the primary use case of collaborative filtering in recommendation systems? 1 2 5

(A) Predicting future stock market prices (B) Recommending movies or products based on user preferences
(C) Analyzing sentiment in social media data (D) Detecting fraud in financial transactions

PART - B (5 × 4 = 20 Marks)

Answer any 5 Questions

Marks BL CO

21. Explain what Hadoop Streaming is and how it can be used to perform MapReduce tasks with non-Java programming languages. Provide a step-by-step guide, including the necessary commands and considerations when using Hadoop Streaming. 4 2 1

22. Explain the key components and their roles within the Hadoop ecosystem. Provide a brief overview of how these components work together to enable distributed data processing and storage. 4 3 1

23. Explain the architecture and key features of the Hadoop Distributed File System (HDFS). Discuss the advantages and use cases of HDFS in the context of big data processing. 4 2 2

24. Explain the concept of file-based data structures, focusing on their characteristics, advantages, and common use cases. Provide examples of file-based data structures to illustrate their practical applications. 4 3 2

25. Explain the key components and phases that constitute the anatomy of a MapReduce job. Describe the roles and interactions of these components, and illustrate the flow of data through the MapReduce framework using a practical example. 4 2 3

26. Compare and contrast Hive with traditional databases. outline the advantages of Hive and its architectural significance. 4 3 4

27. Explain the concept of unsupervised learning in machine learning. Describe the key characteristics of unsupervised learning algorithms and provide an example of a real-world application where unsupervised learning is applied. 4 2 5

PART - C (5 × 12 = 60 Marks)

Answer **all** Questions

Marks BL CO

28. (a) (i) Examine IBM's Big Data Strategy comprehensively. Describe the key components and initiatives that constitute IBM's approach to Big Data. 12 4 1
(ii) Detail on the use of Infosphere Big Insights and its components.

(OR)

(b) i) Discuss the concept of data partitioning and shuffling in the context of Hadoop's MapReduce. Explain their significance and potential bottlenecks in large-scale data analysis. (6 marks)
ii) Compare and contrast Hadoop with alternative big data processing frameworks like Apache Spark and Apache Flink. Highlight the key differences, advantages, and scenarios where each framework is most suitable for analyzing data at scale. (6 marks)

29. (a) Discuss the different ways in which data is ingested into Hadoop with Flume and Sqoop. 12 4 2

(OR)

(b) Elaborate on core components of HDFS detail how serialization and compression is done in Hadoop.

30. (a) With char code snippets, and illustration detail with an example on the working of a map reduce job. 12 3 3

(OR)

(b) Discuss on the map reduce types and input output formats.

31. (a) Detail with examples and codesnippets on table creation, data insertion and querying with HiveQL code snippets. 12 4 4

(OR)

(b) (i) Explore the different execution modes of Pig in more detail. Discuss the advantages, limitations, and use cases for each mode, emphasizing scenarios where one mode might be preferred over the others.
(ii) List and detail with examples of 3 pig latin commands.

32. (a) Compare and contrast on the RDBMS and Hbase databases. Detail on the core architectural concepts of Hbase. 12 3 5

(OR)

(b) Explain the concept of Big data analytics with BigR. Discuss its key components advantages and limitations. provide example of real-world applications where BigR can be effectively utilized.

* * * * *