Reg. No															

B.Tech DEGREE EXAMINATION, NOVEMBER 2023

Fifth & Seventh Semester

18CSE391T - BIG DATA TOOLS AND TECHNIQUES

(For the candidates admitted during the academic year 2020 - 2021 & 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

Marks BL CO

1. Applications and services that run on a distributed network using virtualized resources is known as __________ 1 1 1
(A) Parallel computing (B) Soft computing
(C) Distributed computing (D) Cloud computing

2. Which of the following architectural standards is working with cloud computing industry? 1 1 2
(A) Web-application frameworks (B) Service-oriented architecture
(C) Standardized Web services (D) Adhoc Architecture

3. ________ is the most popular high-level Java API in Hadoop ecosystem. 1 1 2
(A) Scalding (B) HCatalog
(C) Cascalog (D) Cascading

4. __________ is general-purpose computing model and runtime system for distributed data analytics. 1 1 2
(A) MapReduce (B) Drill
(C) Oozie (D) JSON

5. The Pig Latin scripting language is not only a higher-level data flow language but also has operators similar to __________ 1 1 3
(A) SQL (B) JSON
(C) XML (D) Flume

6. What is the primary goal of distributed computing in the context of Big Data? 1 1 3
(A) To centralize all data processing (B) To increase the size of individual servers
(C) To distribute data and processing across multiple nodes (D) To minimize data storage

7. What is the role of DataNodes in HDFS? 1 2 3
(A) They manage the metadata of the file system. (B) They store and manage the actual data blocks.
(C) They schedule and monitor MapReduce jobs. (D) They control access to HDFS.

8. Which HDFS feature ensures data redundancy and fault tolerance by replicating data across multiple DataNodes? 1 2 3
(A) HDFS Compression (B) HDFS Encryption
(C) HDFS Replication (D) HDFS Partitioning

9. Which component of YARN is responsible for managing and allocating cluster resources to applications? 1 1 3
(A) Node Manager (B) Resource Manager
(C) Application Master (D) Data Node

10. ________ is a platform for constructing data flows for Extract, Transform, and Load (ETL) processing and analysis of large datasets. 1 1 4
(A) Pig Latin (B) Oozie
(C) Pig (D) Hive

11. Which of the following describes the "leader" role in a ZooKeeper ensemble? 1 2 3
(A) The node with the slowest processing speed (B) The node responsible for coordination and accepting client requests
(C) A type of znode in ZooKeeper (D) A role that stores data exclusively

12. What is the primary purpose of Apache Flume in the Hadoop ecosystem? 1 2 4
(A) Data processing (B) Data storage
(C) Data collection and ingestion (D) Data analysis

13. Which of the following is NOT a common source type in Apache Flume? 1 1 4
(A) Apache Kafka (B) Apache HBase
(C) Log files (D) Twitter data

14. In Apache Flume, what is a "sink"? 1 1 4
(A) A source of data (B) A type of data transformation
(C) A destination where data is sent (D) A data processing engine

15. What is the primary programming language for developing applications in Apache Spark? 1 1 4
(A) Python (B) Java
(C) C++ (D) Ruby

16. Which of the following is a key feature of Apache Spark that distinguishes it from traditional MapReduce? 1 2 5
(A) Real-time data processing (B) Centralized data storage
(C) Data compression (D) Batch processing only

17. Which of the following best describes MongoDB's data model? 1 2 5
(A) Relational (B) Document-oriented
(C) Key-value (D) Graph-based

18. What is the primary query language used in MongoDB? 1 1 6
(A) SQL (Structured Query Language) (B) NoSQL Query Language (NQL)
(C) JSON Query Language (JQL) (D) BSON Query Language (BQL)

19. What is Tableau primarily used for in the context of data analysis? 1 1 6
(A) Data storage and backup (B) Data collection
(C) Data visualization and analytics (D) Data encryption

20. Which technology is commonly used in Enterprise Data Science for handling and analyzing large datasets? 1 2 6
(A) Virtual reality (B) Augmented reality
(C) Big Data platforms (e.g., Hadoop, Spark) (D) Blockchain

PART - B (5 × 4 = 20 Marks)

Answer **any 5** Questions

Marks BL CO

21. Explain the concept of fault tolerance in distributed computing and its importance in big data processing. How do systems like Hadoop handle faults to ensure the reliability of data processing jobs? 4 2 1

22. Explain the role of Hadoop Distributed File System (HDFS) in the Hadoop ecosystem. How does it provide fault tolerance and high availability for big data storage? 4 2 2

23.	Explain the concept of Hadoop Streaming in MapReduce. How does it enable the integration of non-Java programs into the Hadoop ecosystem?	4	2	2
24.	Explain the purpose and significance of YARN in the Hadoop ecosystem. How does it address resource management challenges in big data processing?	4	2	3
25.	Explain the primary purpose of Sqoop in the Hadoop ecosystem. How does Sqoop facilitate the efficient and scalable transfer of data between Hadoop and relational databases?	4	2	4
26.	Discuss the role of Spark's cluster manager and the different cluster manager options available for Spark applications.	4	2	5
27.	Explain the importance of data visualization in the field of big data analytics. How does effective data visualization enhance the understanding of trends, patterns, and insights within large datasets?	4	2	6

PART - C (5 × 12 = 60 Marks)

Answer **all** Questions

		Marks	BL	CO
28.	(a) (i) Justify the connectivity of big data analytics with big data mining. (ii) Explain in detail the Technical elements of the Big Data platform. **(OR)** (b) (i) Discuss Hadoop Ecosystem and the tools involved in each layer. (ii) Discuss the core modules of Hadoop.	12	2	1
29.	(a) Data replication is a fundamental mechanism in HDFS to ensure data reliability and fault tolerance. Describe how HDFS handles data replication across multiple nodes within a cluster. Discuss the factors that influence the replication factor and how it impacts data durability and availability. Provide examples of scenarios where adjusting the replication factor might be necessary and explain the trade-offs involved. Additionally, address strategies for data recovery in case of node failures or data corruption in HDFS. Conclude with recommendations for optimizing data replication in HDFS for various use cases. **(OR)** (b) Discuss the concept of data partitioning, shuffling, and combiners in MapReduce, and explain how they contribute to efficient data processing in distributed clusters. Provide real-world examples or use cases where organizations have successfully applied these strategies to process massive volumes of data. Additionally, address fault tolerance mechanisms in MapReduce and how they ensure job completion in the presence of node failures. Justify the commendations for designing scalable and fault-tolerant MapReduce implementations for big data processing.	12	3	3
30.	(a) Explain the role of Pig in Hadoop file system. Discuss the steps involved in conversion of user end request to HDFS commands with suitable diagrams. **(OR)** (b) Describe the mechanisms and strategies employed by Apache Flume to ensure the reliable and fault-tolerant transfer of log data. Discuss the role of Flume agents, channels, and sinks in achieving these objectives. Provide examples of how Flume's reliability features have helped organizations maintain data integrity and minimize data loss in the face of various failures. Additionally, address scalability considerations and how Flume can handle increased data volume and sources efficiently. Justify the design resilient log data ingestion pipelines with Apache Flume.	12	3	4

31. (a) Discuss the various data preprocessing steps that are essential before conducting data mining tasks with NoSQL databases. Explain how these steps help in data cleaning, transformation, reduction, and integration. Provide examples or case studies illustrating how effective data preprocessing has improved the accuracy and efficiency of data mining activities in big data projects that use NoSQL databases. Conclude with recommendations for best practices in data preprocessing for big data mining with NoSQL. 12 3 5

(OR)

(b) Explain the principles and considerations for data modeling in MongoDB, particularly when dealing with large and complex datasets. Discuss the differences between schema-less and schema-on-read approaches and provide examples of when each approach is more suitable in a big data context. Additionally, address indexing strategies, sharding, and data partitioning techniques specific to MongoDB and how they impact query performance. Justify the effective data modeling practices in MongoDB for big data applications.

32. (a) Discuss the principles and best practices that should guide the design and implementation of data visualizations in a big data context. Address the issues such as data selection, visualization types, interactivity, and storytelling. Provide examples of how adhering to these principles has led to successful data visualization projects in real-world scenarios. Additionally, discuss the ethical considerations related to data visualization in the context of big data, including potential biases and misinterpretations. 12 3 6

(OR)

(b) Data governance is a critical aspect of successful enterprise data science initiatives, particularly in the context of big data. Explain the key principles and practices of effective data governance for big data projects. Discuss how data governance ensures data quality, security, and compliance in large-scale data environments. Provide examples of organizations that have successfully implemented robust data governance frameworks and the benefits they have realized

* * * * *