Reg. No.														

B.Tech. DEGREE EXAMINATION, MAY 2024

Sixth Semester

18CSE391T – BIG DATA TOOLS AND TECHNIQUES

(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 & Part - C** should be answered in answer booklet.

Time: 3 hours | Max. Marks: 100

	Marks	BL	CO	PO

PART – A (20 × 1 = 20 Marks)

Answer **ALL** Questions

1. Which daemon is responsible for managing the metadata of files stored in HDFS? | 1 | 1 | 2 | 3

(A) Name node (B) Resource manager

(C) Secondary name node (D) Data node

2. What role does the resource manager play in YARN? | 1 | 2 | 2 | 3

(A) It manages the execution of individual tasks on the cluster (B) It allocates the resources to different applications and tracks resource usage

(C) It stores the metadata of the distributed file system (D) It monitors the health of worker nodes in the cluster

3. How does YARN facilitate multitenancy in Hadoop cluster? | 1 | 2 | 2 | 3

(A) By providing strict isolation between different applications (B) By limiting the number of applications that can run concurrently

(C) By sharing the resources dynamically among multiple applications (D) By segregating clusters for different user groups

4. Which of the following represents the current order of phases in a map reduce job? | 1 | 1 | 2 | 3

(A) Map, reduce, shuffle (B) Shuffle, map, reduce

(C) Map, shuffle, reduce (D) Reduce, map, shuffle

5. Which core module of Hadoop is responsible for managing and processing large-scale, real time data streams? | 1 | 1 | 1 | 3

(A) HDFS (B) YARN

(C) SPARK (D) HBASE

6. Which core module of Hadoop provides faults tolerance and resource management for distributing applications? | 1 | 1 | 1 | 3

(A) HDFS (B) YARN

(C) HIVE (D) OOZIE

7. What is Hadoop? 1 2 1 3

(A) A programming language
(B) A distributed file system and data processing framework
(C) A relational database management system
(D) A cloud computing platform

8. The core module of Hadoop designed for scalable, high performance batch processing is _____. 1 1 1 3

(A) Pig
(B) HDFS
(C) Map reduce
(D) Hbase

9. How does zookeeper ensure fault tolerance 1 2 3 3

(A) By replicating data across multiple nodes
(B) By compressing data to reduce storage space
(C) By encrypting data for security
(D) By distributing data evenly across the cluster

10. In sqoop, what is a connector? 1 2 3 3

(A) The transformation applied to incoming data
(B) The source of data to be ingested
(C) The destination where data is sent
(D) A plugin that provides connectivity to a specific database or data source

11. Which NoSQL databases is known for its ability to handle distributed and fault-tolerance data storage? 1 1 4 2

(A) IBMDB2
(B) Cassandra
(C) Redis
(D) Couchbase

12. In HBase, what is a region? 1 1 4 2

(A) A distributed file system
(B) A data structure for storing data records
(C) A logical partition of the data stored in a table
(D) A plugin that provides connectivity to a specific database / data source

13. How does apache HBase achieve fault tolerance? 1 1 4 3

(A) By replicating data across multiple nodes
(B) By compressing data to reduce storage space
(C) By encrypting data for security
(D) By maintaining multiple copies of the same data blocks in different region servers

14. Which of the following actions can oozie coordinates and manage? 1 1 4 3

(A) Data integration
(B) Realtime data analysis
(C) Hadoop job execution
(D) Configuration management

15. What is the primary advantage of using Pig Latin over traditional map reduce programming 1 1 3 3

(A) Pig Latin offers better fault tolerance
(B) Pig Latin allows for faster development of data processing tasks
(C) Pig Latin provides stronger consistency guarantees
(D) Pig Latin is designed for real-time data processing

16. What is the primary purpose for sqoop import command? 1 1 3 3

(A) To export data from Hadoop to a relational database
(B) To import data from a relational database to Hadoop
(C) To transform data between sources and sinks
(D) To provide fault tolerance and durability

17. Which python library is commonly used to create publication quality static plots? 1 1 5 3

(A) Matplotlib
(B) Seaborn
(C) Plotly
(D) Bokeh

18. Which product of Oracles Exa family is a complementary solution for BI workloads? 1 1 5 3

(A) Exalytics
(B) Exologic
(C) Exadata
(D) Timer ten

19. What is SLA? 1 1 5 3

(A) Service-level agreement
(B) Source-level agreement
(C) Solution level agreement
(D) Service-lease agreement

20. Which is an extension of the 'q' programming language? 1 1 5 3

(A) Kdb+
(B) Green plum
(C) SAP hana
(D) Mango DB

	Marks	BL	CO	PO

PART – B (5 × 4 = 20 Marks)
Answer **ANY FIVE** Questions

21. Discuss the role of mappers and reducers in data transformation and aggregation. 4 3 1 3

22. Compare and contrast Hadoop's YARN(Yet Another Resource Negotiator) with the classic Hadoop Map Reduce Framework. 4 3 2 3

23. Explain briefly the key features of HDFS, such as replication, fault tolerance and data locality. 4 3 2 3

24. Mention the technical elements of big data platform. 4 3 1 3

25. Briefly discuss flume architecture. 4 3 3 3

26. What is oozie and its role in HDFS? 4 3 4 3

27. Differentiate discrete and continuous data with examples. Compare and differentiate bar plots and histogram. 4 3 5 3

	Marks	BL	CO	PO

PART – C (5 × 12 = 60 Marks)
Answer **ALL** Questions

28. a. Explain in detail about the core modules of Hadoop. 12 3 1 3

(OR)

b. Briefly describe various analytics toolkit associated with big data mining and explain the various components of the analytical tool kit. 12 3 1 3

29. a.	Explain the key differences between "Map" and "Reduce" phases in a map reduce job with suitable example.	12	4	2	2
	(OR)				
b.	Explain the concepts of Resource managers and node managers in YARN. Describe how they communicate and collaborate to schedule and monitor application execution.	12	4	2	3
30. a.	Illustrate the steps involved in setting up Hadoop cluster and the detail the working of core components of Hadoop.	12	4	3	3
	(OR)				
b.	What is Pig Latin and how does it relate the Apache pig? Discuss the purpose and basic syntax of Pig Latin. Explain in brief DISTINCT, DUMP, STORE and ORDER by commands used in pig Latin.	12	4	3	3
31. a.	Discuss on the advantage of SPARK over similar frameworks. Detail on the data structures used by SPARK to ensure fault tolerance.	12	3	4	2
	(OR)				
b.	How does NoSQL handle unstructured or semi-structured data compared to traditional SQL databases? Provide example with code snippets for MongoDB DML and DDL statements.	12	3	4	2
32. a.	List and discuss the challenges associated with big data analytics and the road map to enterprise analytical success.	12	4	5	3
	(OR)				
b.	Explore the ecosystem sourcing R and python with respect to data visualization and evaluate the maturity and versatility of each ecosystem in supporting diverse visualization needs.	12	4	5	3

* * * * *