Reg. No															

B.Tech DEGREE EXAMINATION, NOVEMBER 2023

Seventh Semester

18CSE333J - BIG DATA TOOLS AND TECHNIQUES FOR BLOCKCHAIN

(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.	What is Big Data? (A) Data with large fonts (B) Data with complex formats (C) Extremely large datasets (D) Data that is difficult to understand	1	1	1
2.	Which of the following is NOT a characteristic of Big Data? (A) Volume (B) Velocity (C) Variety (D) Validity	1	1	1
3.	What is the primary goal of Hadoop in Big Data processing? (A) Real-time data analytics (B) Efficient storage and processing of large datasets (C) Data visualization (D) Data compression	1	2	1
4.	Which programming language is commonly used for writing MapReduce jobs in Hadoop? (A) Java (B) Python (C) C++ (D) Ruby	1	1	1
5.	What does HDFS stand for in the context of Hadoop? (A) Hadoop Database Storage File System (B) Highly Distributed File Storage (C) Hadoop Distributed File System (D) High-Density File Storage	1	1	2
6.	In HDFS, what is the default replication factor? (A) 1 (B) 2 (C) 3 (D) 4	1	1	2
7.	Which component of HDFS is responsible for storing actual data blocks? (A) ResourceManager (B) NameNode (C) DataNode (D) Secondary NameNode	1	2	2
8.	What is the purpose of the "Secondary NameNode" in HDFS? (A) To take over if the primary NameNode fails (B) To provide backup storage for data (C) To periodically merge namespace and edit logs (D) To manage the ResourceManager	1	2	2
9.	What is speculative execution in the context of MapReduce? (A) A backup mechanism for Map and Reduce tasks (B) The process of running multiple instances of the same task to improve job completion time (C) A technique for compressing intermediate data (D) The process of replicating data across multiple DataNodes	1	3	3

10. What is the purpose of the JobTracker in a Hadoop MapReduce cluster? 1 2 3
(A) To manage the execution of Map and Reduce tasks
(B) To store and manage the input data for jobs
(C) To control the distribution of data across DataNodes
(D) To schedule jobs based on user priorities

11. In MapReduce, what is the purpose of a Combiner function? 1 4 3
(A) To combine output from multiple Reducers into a single result
(B) To combine the output of the Mapper and Reducer
(C) To perform pre-processing on input data before mapping
(D) To aggregate data at the Mapper stage

12. What is speculative execution in the context of MapReduce? 1 5 3
(A) A mechanism for handling speculative tasks in a cluster
(B) The process of running multiple instances of the same task to improve job completion time
(C) A technique for handling speculative failures in Hadoop
(D) The process of replicating data across multiple DataNodes

13. Which component of the Hadoop ecosystem is responsible for distributed storage and processing of data? 1 2 4
(A) Hive
(B) HDFS (Hadoop Distributed File System)
(C) Pig
(D) Spark

14. What is the primary function of Apache Hive in the Hadoop ecosystem? 1 2 4
(A) Real-time data processing
(B) Batch processing of structured data using SQL-like queries
(C) Stream processing of data
(D) Data visualization

15. Which Hadoop ecosystem component is designed for processing and analyzing large-scale datasets using a batch processing model? 1 3 4
(A) Spark
(B) HBase
(C) YARN (Yet Another Resource Negotiator)
(D) Oozie

16. What does Apache Pig offer in the Hadoop ecosystem? 1 2 4
(A) Real-time data processing
(B) A high-speed data transfer protocol
(C) A platform for building MapReduce jobs using a scripting language
(D) A distributed NoSQL database

17. What is the "cold start problem" in collaborative filtering? 1 4 5
(A) The challenge of dealing with cold weather during data analysis
(B) The issue of starting a machine learning model from scratch
(C) A problem where new users or items have limited data for recommendations
(D) The difficulty of deploying machine learning models to production

18. Which machine learning algorithm is commonly used for natural language processing (NLP) tasks like text classification and sentiment analysis? 1 3 5
(A) k-Nearest Neighbors (k-NN)
(B) Support Vector Machines (SVM)
(C) Recurrent Neural Networks (RNN)
(D) Principal Component Analysis (PCA)

19. What is the primary difference between content-based filtering and collaborative filtering in recommendation systems? 1 4 5

(A) Content-based filtering relies on user preferences, while collaborative filtering relies on item characteristics.

(B) Content-based filtering uses deep learning models, while collaborative filtering uses decision trees.

(C) Content-based filtering is unsupervised, while collaborative filtering is supervised.

(D) Content-based filtering focuses on cold start problems, while collaborative filtering does not.

20. In collaborative filtering, what does "user-user" collaborative filtering involve? 1 4 5

(A) Comparing users based on their personal characteristics

(B) Finding similarities between users based on their behavior and preferences

(C) Collaborating with users to build machine learning models

(D) Analyzing user-generated content on social media platforms

PART - B (5 × 4 = 20 Marks) Marks BL CO

Answer **any 5** Questions

21. Explain the three primary types of digital data, providing examples for each type. Also, discuss the significance of understanding these data types in the context of data management. 4 2 1

22. Explain the process of analyzing large-scale data using Hadoop MapReduce. Provide a step-by-step guide, including the key components involved, and explain their roles. Use a practical example to illustrate the MapReduce process. 4 2 1

23. Describe the process of data replication in HDFS, including its purpose, benefits, and how it ensures fault tolerance. Discuss the trade-offs involved in choosing the replication factor for data stored in HDFS. 4 2 2

24. Explain the different Hadoop file system interfaces, namely HDFS (Hadoop Distributed File System) and Hadoop Common, highlighting their roles, characteristics, and use cases within the Hadoop ecosystem. 4 2 2

25. Explain the MapReduce programming model, its core components, and the steps involved in processing data using MapReduce. Provide a practical example to illustrate the MapReduce process. 4 2 3

26. Explain what Apache Pig is, its key features, and how it simplifies data processing tasks in the context of Hadoop. Provide a practical example of how Pig Latin scripts are used to process data. 4 4 4

27. Explain the concept of supervised learning in machine learning. Describe the key components of supervised learning, including the role of training data, features, labels, and the process of model training. Provide an example of a real-world application of supervised learning. 4 2 5

PART - C (5 × 12 = 60 Marks) Marks BL CO

Answer **all** Questions

28.	(a) Examine the concept of Big Data Analytics comprehensively. Describe the key components and challenges of Big Data Analytics. Discuss the technologies and techniques commonly used for processing and extracting insights from large datasets. Provide examples of real-world applications where Big Data Analytics has made a significant impact.	12	4	1
	(OR)			
	(b) i) Explain the Hadoop ecosystem components that are essential for data analysis. Discuss the roles of Hadoop Distributed File System (HDFS), MapReduce, YARN, and Hive in the context of analyzing large datasets. (6 marks) ii) Describe the process of setting up a Hadoop cluster for data analysis. Include the hardware requirements, software configuration, and any best practices for ensuring scalability, fault tolerance, and performance. (6 marks)			
29.	(a) Examine the design principles and architecture of HDFS (Hadoop Distributed File System) comprehensively. Describe the key components, data replication strategies, and fault tolerance mechanisms that constitute the design of HDFS. Discuss the advantages and trade-offs of HDFS in handling big data. Provide real-world examples where HDFS has been a valuable asset in data management and processing.	12	5	2
	(OR)			
	(b) i) Explain the different file system interfaces in Hadoop, focusing on Hadoop Distributed File System (HDFS) and its architecture. (6 marks) ii) Compare and contrast the HDFS with other distributed file systems like Amazon S3 and Azure Data Lake Storage. Highlight the advantages and disadvantages of using each file system for big data workloads. (6 marks)			
30.	(a) Examine the key functions and principles of the MapReduce programming model in the context of distributed data processing. Describe in detail how MapReduce accomplishes data processing tasks, including data input, mapping, shuffling, sorting, reducing, and output. Provide real-world examples to illustrate the MapReduce functions in action.	12	4	3
	(OR)			
	(b) i) Describe the different types of MapReduce programming models, including classic MapReduce, YARN-based MapReduce, and Spark's MapReduce-style transformations. Explain the key differences and use cases for each type. (6 marks) ii) Discuss the merits and demerits of using MapReduce as a data processing paradigm in distributed computing. Provide examples of scenarios where MapReduce is particularly suitable and where it may not be the best choice. (6 marks)			
31.	(a) Compare and contrast Apache Pig, a high-level data processing platform, with traditional relational databases. Examine the key characteristics, use cases, and advantages of Pig in handling large-scale data processing tasks. Describe the scenarios where Pig is preferred over databases and vice versa. Provide examples to illustrate the differences and similarities between Pig and databases.	12	5	4
	(OR)			
	(b) i) Explain the key components of the Hadoop ecosystem, including HDFS, MapReduce, Hive, Pig, HBase, and Spark. Describe the roles and use cases of each component within the Hadoop ecosystem and how they work together to enable big data processing. (6 marks) ii) Dive deeper into Pig's role within the Hadoop ecosystem. Discuss the advantages of using Pig for data processing, its data model, and the different execution modes available in Pig. Provide examples of scenarios where Pig is a suitable choice for data processing. (6 marks)			

32. (a) Examine the role and significance of User-Defined Functions (UDFs) in data analytics. Describe the types of UDFs commonly used in data analytics and their applications. Discuss the advantages and challenges of implementing UDFs in data analysis workflows. Provide examples and scenarios to illustrate the use and impact of UDFs in data analytics. 12 5 5

(OR)

(b) i) Explain the key components of the Hadoop ecosystem, including HDFS, MapReduce, Hive, Pig, HBase, and Spark. Describe the roles and use cases of each component within the Hadoop ecosystem and how they work together to enable big data processing. (6 marks)

ii) Dive deeper into Pig's role within the Hadoop ecosystem. Discuss the advantages of using Pig for data processing, its data model; and the different execution modes available in Pig. Provide examples of scenarios where Pig is a suitable choice for data processing. (6 marks)

* * * * *