Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

B.Tech Degree S5 (S,FE) Examinations June 2026 (2019 Scheme)

Course Code: CDT307
Course Name: BIG DATA PROCESSING

Max. Marks: 100 Duration: 3 Hours

PART A

(Answer all questions; each question carries 3 marks) Marks

1 Select either RDBMS or Hadoop for analysing a large dataset of social media posts and justify your choice. 3

2 Given a Data Frame student. Write the output for student[2] and student[[2]] 3

	Roll	Mark1	Mark2
1	1	20	20
2	2	30	30
3	3	40	40

3 Discuss the significance of Hadoop-specific file types, such as SequenceFile or Avro. In what scenarios would you choose to use these file types over traditional file formats? 3

4 Describe the compaction process in HBase. How does it affect read and write performance? 3

5 What is a Combiner in MapReduce, and how does it enhance performance? Provide an example of its use. 3

6 Compare MapReduce with traditional data processing techniques. What are the key differences and advantages? 3

7 Explain the concept of the filter operation in stream processing. How does it contribute to the efficiency of data processing in streaming applications? Provide an example of a use case where filtering is essential. 3

8 Write a Hive query to add a new column description of type STRING to an existing table product. 3

9 Enumerate the distinctive characteristics that set Apache Spark apart from other big data processing solutions. 3

10 Explain the use of DUMP in Piglatin 3

PART B

(Answer one full question from each module, each question carries 14 marks)

Module -1

11	a)	Describe the significance of the Hadoop ecosystem in the context of big data processing and analytics.	10
	b)	Given a scenario where a Hadoop cluster is facing performance issues due to a single NameNode, outline the steps to implement HDFS Federation in the existing architecture. What changes would need to be made?	4
12	a)	A CSV file named sales_data.csv contains columns for Product, Sales, and Region. Write an R script to load this data and filter out all records where Sales exceed $1000.	7
	b)	Create a function named square_and_sum that takes a numeric vector as input, squares each element, and then returns the sum of the squared values. Apply this function to the vector [2, 3, 4] and display the result.	7

Module -2

13	a)	Analyse the design principles of HDFS. How do these principles impact data availability and fault tolerance in a distributed environment?	7
	b)	Discuss the differences between HBase and HDFS in terms of data access patterns. In which scenarios would you prefer to use HBase over HDFS for data storage?	7
14	a)	Describe the HBase data model, focusing on tables, column families, and rows. How does this model differ from traditional relational databases, and what advantages does it offer?	4
	b)	Identify and briefly describe the key components of HBase architecture. How do they work together to manage data?	10

Module -3

15	a)	Explain the detailed steps involved in the execution pipeline of a MapReduce job. How does each phase contribute to the overall process?	10
	b)	Discuss the best practices for designing efficient MapReduce jobs. What considerations should be made regarding data locality and resource allocation?	4
16	a)	Explain the process of creating inverted indexes using MapReduce. What steps are required, and how do they improve search efficiency?	7
	b)	Describe the MapReduce programming model, including its components and workflows. How does this model facilitate parallel processing?	7

Module -4

17	a)	Differentiate adhoc Queries and standalone queries with example	4
	b)	Explain the significance of sampling in stream data model. Describe other operation .	
18	a)	Write down the Hive queries for the following (i)Create a table named user_interaction with fields interaction_time, user_id, interaction_type, content_id, and session_id. Add a comment "Session ID of the user" to the session_id field. Define partitions based on interaction_date and platform. The cluster must be defined based on user_id with 16 buckets. (ii) Get all user interactions recorded in May 2024 for interactions of type 'click' on the platform 'mobile'. (iii) Display all clusters and partitions in the user_interaction table. (iv) Delete a partition from the user_interaction table.	8
	b)	Differentiate external tables and managed tables	6

Module -5

19	a)	Explain the structure and purpose of a Pig Latin script. Provide a simple example script that loads a dataset, filters it, and stores the results.	6
	b)	Compare and contrast the characteristics of Atoms, Tuples, Bags, and Maps in Pig, highlighting their individual roles in data representation and processing.	8
20	a)	Describe the fundamental building blocks within the architecture of Apache Spark	10
	b)	Describe the various operations that can be performed on RDDs in Apache Spark.	4

***