

Reg No.: _____

Name: _____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

B.Tech Degree S5 (R,S) Examination November 2025 (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 | Explain the role of Hadoop in solving business problems? | 3 |
| 2 | Write a R program to create a list of numeric vectors, then use lapply() to find the sum of each element in the list and display the result. | 3 |
| 3 | Explain the role of rack awareness in HDFS replication and discuss how it affects both read and write performance within a distributed cluster? | 3 |
| 4 | Explain the roles of region servers and the HMaster in HBase's data management architecture? | 3 |
| 5 | Discuss the role of mappers and reducers in MapReduce Framework. | 3 |
| 6 | Write a MapReduce workflow for processing millions of images using a face-recognition algorithm that identifies a set of desired features and produces recognition results. | 3 |
| 7 | Explain general sampling problem? | 3 |
| 8 | Discuss collection datatypes in Hive with examples. | 3 |
| 9 | List any three Pig built-in functions under Eval() category? | 3 |
| 10 | Explain the concept of lazy evaluation in Apache Spark? | 3 |

PART B*(Answer one full question from each module, each question carries 14 marks)***Module -1**

- 11 a) With the help of a neat diagram, explain the core components of the Hadoop ecosystem and how they work together for efficient big data processing? 10
- b) Describe the key components and techniques used to handle the challenges of Big Data. 4
- 12 a) Write a R program to create a data frame containing details of four employees (name, age, and salary). Filter the data frame to display only the employees with a salary greater than 50,000, and calculate the average salary of these employees. 8
- b) Explain how R handles file input and output operations for different file formats. 6

Module -2

- 13 a) With help of a diagram, illustrate the architecture of HDFS and explain how it manages large files across a Hadoop cluster? 10
- b) Draw the record-based and block-based layout of SequenceFile in Hadoop and explain their structure. 4
- 14 a) With a neat diagram, explain the use of HBase auto-sharding and distribution approach of tables and column families. 10
- b) Summarize the role of the (.META.) table in HBase and explain how it helps resolve a key/table pair to the appropriate region server? 4

Module -3

- 15 a) Describe the execution pipeline of a MapReduce job with the help of a diagram. 10
- b) Summarize the role of the JobTracker and TaskTracker in the MapReduce framework. 4
- 16 a) Examine the approach for solving the road enrichment problem using MapReduce. 8
- b) Design a MapReduce job for the calculation of inverted indexes in a data processing application. 6

Module -4

- 17 a) Explain Flajolet-Martin Algorithm for counting distinct elements in a stream. 8
- Suppose the data stream consists of the integers 4, 2, 5, 9, 1, 6, 3, 7, 2, 1. Determine

the tail length for each stream element and the resulting estimate of the number of distinct elements if the hash function is $h(x) = x + 6 \bmod 32$ (treat the result as a 5-bit binary integer).

- b) Summarize the architectural differences between handling continuous data streams and traditional database queries. 6
- 18 a) Write down Hive queries for the following: 8
- (i) Create a table `user_activity` with fields `activity_time`, `user_id`, `action`, `device`, and `location`. Add a comment "User's device information" to the device field. Define partitions based on region and date. The table should be clustered by `user_id` into 16 buckets.
 - (ii) Retrieve all activities performed by users from the 'IN' region in the month of 2020-02-23 where the action is "login".
 - (iii) Show all partitions in the `user_activity` table.
 - (iv) Delete a partition from the `user_activity` table for the region 'IN' and the date 2022-02-15.
- b) Write the syntax of creating managed tables and external tables in Hive and mention the key differences between them. . 6

Module -5

- 19 a) Explain the architecture of Apache Spark, describing the role of each component in the data processing framework. 10
- b) Explain the Apache Pig data model, highlighting its various data types and structures. 4
- 20 a) Describe the various operations that can be performed on RDDs in Apache Spark. 8
- b) Explain with example, how user defined functions are implemented in Pig. 6
