**23CSPC402**

USN | | | | | | | | | | |

# MANGALORE INSTITUTE OF TECHNOLOGY & ENGINEERING

(A Unit of Rajalaxmi Education Trust®, Mangalore)
Autonomous Institute affiliated to VTU, Belagavi, Approved by AICTE, New Delhi
Accredited by NAAC with A+ Grade & ISO 9001:2015 Certified Institution

## Seventh Semester B.E Degree Examination Model Question Paper

# Big Data Analytics

**Time: 3 Hours (180 Minutes)** **Max. Marks: 100**

***Note: 1. Answer any FIVE full questions, choosing ONE full question from each module.***
***2. M: Marks, L: RBT (Revised Bloom's Taxonomy) level, C: Course outcomes.***

| | | Module -1 | M | L | C |
|---|---|---|---|---|---|
| Q1 | a. | Explain the classification of digital data with suitable examples. | 6 | L2 | CO1 |
| | b. | Explain characteristics of Data and challenges with Big Data | 6 | L2 | CO1 |
| | c. | Explain In-Memory Analytics, In-Database Processing, and Massively Parallel Processing. Write any two differences between parallel and distributed systems with suitable examples. | 8 | L2 | CO1 |
| | | **OR** | | | |
| Q2 | a. | Define Big Data. Explain three V's, the core characteristic of Big Data with suitable examples. | 6 | L2 | CO1 |
| | b. | Explain the advantages of NoSQL and compare SQL, NoSQL, and NewSQL databases. | 6 | L2 | CO1 |
| | c. | Explain Hadoop Ecosystem Components for Data Processing and Data Analysis. | 8 | L2 | CO1 |
| | | **Module- 2** | | | |
| Q3 | a. | Explain the differences between RDBMS and Hadoop with respect to data type, processing, system structure, and processor requirements. | 6 | L2 | CO1 |
| | b. | A client needs to read a file stored in HDFS. Apply the HDFS file-read process and illustrate, with a suitable block diagram, how the client obtains the block locations from the NameNode and reads the file blocks from the appropriate DataNodes. Explain the sequence of operations involved. | 7 | L3 | CO2 |
| | c. | Develop a Java MapReduce program to find and count the occurrences of similar words in an input text file. | 7 | L3 | CO2 |
| | | **OR** | | | |
| Q4 | a. | Briefly explain the roles of the Combiner and Partitioner in Hadoop MapReduce. How do you compress the output file? | 6 | L2 | CO1 |
| | b. | A WordCount job is to be processed using MapReduce. Apply the MapReduce processing model to trace, with a suitable block diagram, how the input data passes through the Mapper, Combiner, Partitioner, Shuffle and Sort, and Reducer to produce the final output. | 7 | L3 | CO2 |
| | c. | A client needs to store a 192 MB file in HDFS using the default block size (64MB) and replication factor. Apply the HDFS architecture to illustrate, with a suitable block diagram, how the Client, NameNode, and DataNodes work together to store the file. | 7 | L3 | CO2 |
| | | **Module – 3** | | | |
| Q5 | a. | Consider a MongoDB collection named Student containing the fields student_id, name, department, and marks. Assume that the collection contains five student documents. Write MongoDB queries to:<br>i) Insert a new student document.<br>ii) Find all students belonging to the CSE department. | 10 | L3 | CO3 |

| | | | | | |
|---|---|---|---|---|---|
| | | iii) Update the marks of a specified student.<br>iv) Sort the students based on marks in descending order.<br>v) Display the first three documents. | | | |
| | b. | Consider a Cassandra keyspace named CollegeDB. Create a student table with the attribute's student_id, name, department, and marks. Using the student records provided below, write appropriate CQL commands to perform CRUD operations:<br>i) Create the Student table.<br>ii) Insert the given student records.<br>iii) Retrieve all student records.<br>iv) Update the marks of student 101 from 85 to 90.<br>v) Delete the record of student 105. | 10 | L3 | CO3 |
| | | **OR** | | | |
| Q6 | a. | Consider an employee collection containing the fields employee_id, name, department, and salary. The collection contains employee details from the CSE, ECE, and ISE departments. Write a MongoDB aggregation query to calculate the department-wise employee count and total salary. Display the department name, number of employees, and total salary for each department. | 10 | L3 | CO3 |
| | b. | Assume a Cassandra keyspace named CollegeDB. Create a student table containing student_id, name, skills, subjects, and marks, where skills is a SET, subjects is a LIST, and marks is a MAP of subject names and marks. Using suitable student data, write CQL commands to:<br>i) Create the table and insert student records.<br>ii) Add and remove an element from the SET collection.<br>iii) Add an element to the LIST collection.<br>iv) Add and update a key-value pair in the MAP collection.<br>v) Retrieve the collection data. | 10 | L3 | CO3 |
| | | **Module – 4** | | | |
| Q7 | a. | Assume two Hive tables, Employee and Department, containing employee and department information. Perform the following operations using HiveQL:<br>i) Create the Employee table with the attributes employee_id, employee_name, department_id, and salary.<br>ii) Create the Department table with the attributes department_id and department_name.<br>iii) Insert suitable records into both the Employee and Department tables.<br>iv) Write a HiveQL query to perform an appropriate JOIN operation between the two tables using department_id.<br>v) Display the employee name, department name, and salary of all employees obtained from the JOIN operation. | 10 | L3 | CO4 |
| | b. | Consider an employee dataset containing the attributes id, name, department, and salary. Using Pig Latin, perform the following operations:<br>**Input file: employee.csv**<br><br>**Id** \| **Name** \| **Department** \| **Salary**<br>101 \| Ravi \| CSE \| 45000<br>102 \| Anita \| ECE \| 42000<br>103 \| Arun \| CSE \| 55000<br>104 \| Priya \| ISE \| 48000<br>105 \| Kiran \| CSE \| 60000<br>106 \| Meena \| CSE \| 52000<br>107 \| Rahul \| ME \| 40000<br>108 \| Sneha \| CSE \| 47000 | 10 | L3 | CO4 |

i) Load the employee data from a CSV file named employee.csv and define the schema with appropriate data types.
ii) Use the FILTER operator to select employees belonging to the CSE department.
iii) Use the FOREACH operator to display only the employee name and salary.
iv) Use the ORDER BY operator to arrange the selected employees in descending order of salary.
v) Use the LIMIT operator to display only the first three employees from the ordered result.

**OR**

Q8 a. Consider a Sales dataset. Using HiveQL, perform the following operations:

**Input file: Sales dataset**

| sale_id | product | amount | year | month |
|---|---|---|---|---|
| 101 | Laptop | 55000 | 2025 | January |
| 102 | Mobile | 25000 | 2025 | January |
| 103 | Printer | 15000 | 2025 | February |
| 104 | Keyboard | 2000 | 2025 | February |
| 105 | Laptop | 60000 | 2026 | January |
| 106 | Mobile | 28000 | 2026 | January |
| 107 | Monitor | 18000 | 2026 | March |
| 108 | Mobile | 1200 | 2026 | March |

10 | L3 | CO4

i) Create a database named SalesDB and select it for use.
ii) Create a Sales table containing sale_id, product, and amount as the data columns, with year and month as partition columns.
iii) Insert suitable sales records into different year and month partitions.
iv) Write a HiveQL query to display all the available partitions of the Sales table.
v) Write a HiveQL query to retrieve sales records from a specified partition, such as the year 2025 and month January.

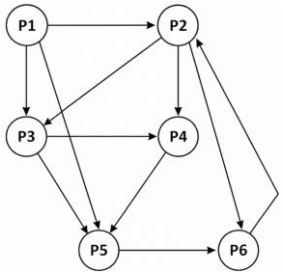
b. Consider an employee dataset containing the attributes id, name, department, and salary. Using Pig Latin, perform the following operations:

**Input file: employee.csv**

| Id | Name | Department | Salary |
|---|---|---|---|
| 101 | Ravi | CSE | 45000 |
| 102 | Anita | CSE | 50000 |
| 103 | Arun | ECE | 40000 |
| 104 | Priya | ECE | 46000 |
| 105 | Kiran | ISE | 55000 |
| 106 | Meena | ISE | 55000 |
| 107 | Rahul | CSE | 60000 |
| 108 | Sneha | ECE | 44000 |
| 109 | Vijay | ISE | 48000 |

10 | L3 | CO4

i) Load the employee data from a CSV file named employee.csv and define the schema with appropriate data types.
ii) Use the GROUP operator to group the employees based on their department.
iii) Use the AVG function to calculate the average salary of employees in each department.
iv) Use the COUNT function to calculate the number of employees in each

| | | | | | |
|---|---|---|---|---|---|
| | | department.<br>v) Display the department name, average salary, and number of employees for each department. | | | |
| **Module – 5** | | | | | |
| Q9 | a. | Describe the key features of Apache Spark and explain how they improve performance compared to traditional Hadoop processing. | 6 | L2 | CO1 |
| | b. | Develop a Spark program to count words in a large text dataset. | 8 | L3 | CO5 |
| | c. | A company collects website log data containing user visits, page views, and session durations. Apply web mining techniques to analyze the logs and identify popular pages and user behavior patterns. | 6 | L3 | CO5 |
| **OR** | | | | | |
| Q10 | a. | Describe how a Web graph can be analyzed in web mining. | 6 | L2 | CO1 |
| | b. | Consider the following web graph, where each node represents a web page and each directed edge represents a hyperlink from the source page to the destination page. Apply the PageRank algorithm using the in-degree as conferring authority to determine the PageRank of each web page. Initialize the PageRank of each page to 1/\|S\|, perform two iterations of the PageRank computation, and rank the pages according to their PageRank values.<br>  <br> | 8 | L3 | CO5 |
| | c. | A news organization wants to analyze thousands of articles to identify trending topics. Apply text mining techniques to extract keywords and categorize the articles automatically. | 6 | L3 | CO5 |

-*****-