Reg. No: 23BCB0015

B/L/TY

Final Assessment Test – November 2025

VIT Vellore Institute of Technology

Course: BCSE402L - Big Data Analytics

Class NBR(s): 1711/1715/1959 Slot: F2+TF2

Time: Three Hours Max. Marks: 100

- KEEPING MOBILE PHONE/ANY ELECTRONIC GADGETS, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE
- DON'T WRITE ANYTHING ON THE QUESTION PAPER

COs	CO Statements
CO1	Recall the characteristics of digital data, data sources, data storage, and the applications of big data in different fields.
CO2	Utilize Hadoop ecosystem tools and Hadoop YARN functions for parallel processing of application tasks.
CO3	Comprehend the Map Reduce programming model and the Map Reduce Daemon framework.
CO4	Apply NoSQL databases for data store management to solve big data problems
CO5	Analyze and evaluate the use of spark stack components with RDDs, ETL built-in functions for handling big data.

BL – Blooms Taxonomy Level (1 – Remember, 2 – Understand, 3 – Apply, 4 – Analyse, 5 – Evaluate, 6 – Create)

Answer ALL Questions

(10 X 10 = 100 Marks)

1. How does Hadoop handle parallel processing of Large Datasets? Also how scalability and parallel processing are achieved in modern Big Data frameworks such as Hadoop and Spark. Discuss the memory management techniques between these platforms. **CO1 BL1**

2. a) Explain how HDFS block size impacts storage efficiency and read/write performance. What considerations determine the optimal block size? **[5] CO2 BL2**

   b) Compare Pig and Hive: when would you prefer Pig over Hive for data transformation tasks? **[5]**

3. An e-commerce platform stores customer ratings of products in the form (ProductID, UserID, Rating). Describe a MapReduce algorithm to efficiently compute the average rating for each product, detailing the Mapper, Combiner, and Reducer functions, and explain the intermediate key-value processing. **CO3 BL3**

4. Demonstrate a multi-step CRUD (Create, Read, Update, and Delete) workflow using CQL for Cassandra for a table tracking employee attendance by writing sample queries for each operation. The employee attendance schema is given below. **CO4 BL4**

Column	Type	Description
employee_id	uuid	Unique employee identifier
date	date	Date of attendance
check_in	timestamp	Time of check-in
check_out	timestamp	Time of check-out
status	text	Attendance status ("Present", "Absent", "On Leave")

OR

10.b) i) Explain the purpose and basic syntax of SPARQL. How does it compare to SQL for querying relational databases? **[4] CO4 BL3**

ii) Write a **SPARQL query** to retrieve all people who work at a specific company from an RDF dataset. **[6]**

⇔⇔⇔ B/L/TY ⇔⇔⇔

5. a) What are transformation and action commands in Spark RDD? Provide two examples for each and describe their effect on data processing. [5] CO5 BL2

b) Describe functional programming basics as supported in Spark. How do functions like map, filter, and reduce enable parallel data processing? [5]

6. What is the significance of estimating moments in stream data? Provide an example where second-order moments might be useful. Explain how frequent item sets are discovered in streaming data. CO5 BL3

7. Explain how ranking algorithms, such as PageRank, work over large graph structures. Present an example of applying PageRank in the context of network organization or social influence. CO5 BL3

8. a) Create a DataFrame from a CSV/JSON file containing employee details. Register it as a temporary SQL table. Write and execute a query to show all records. [2] CO5 BL3

b) Write a Spark SQL query to select the name and age of users from the table where users are older than 30. [2]

c) Find employees with the highest salary per department. [2]

d) Compute the average salary of employees aged under 30, grouped by department. [2]

e) Find departments with more than 5 employees, and the maximum salary in each. [2]

9.a) Explain the **K-Core algorithm** in detail. How can it be used to detect tightly connected communities in a network? CO5 BL4

OR

9.b) Explain and Compare **Lossy Counting** and **Sticky Sampling** algorithms for identifying frequent items. CO5 BL4

10.a) Given a collection orders with documents like: CO4 BL3

```
{
 _id: ObjectId("..."),
 customerId: ObjectId("..."),
 items: [
  { product: "Laptop", quantity: 1, price: 1200 },
  { product: "Mouse", quantity: 2, price: 25 }
 ],
 orderDate: ISODate("2025-10-25T09:00:00Z"),
 status: "shipped"
}
```

i) Write an aggregation query to calculate the total revenue per customer. [3]

ii) Find the top 3 products that generated the highest total revenue across all orders. [3]

iii) Calculate the average order value per month in 2025. [4]