

**VIT**

Vellore Institute of Technology

REG.NO.: 2206E0596

SLOT: A2 +TA2

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2025-2026

Programme Name & Branch : B.Tech. CSE
 Course Code and Course Name : BCSE406L - NoSQL Databases
 Faculty Name(s) : Vani Rajasekar, Anandita Iyer
 Class Number(s) : 1969, 1971
 Date of Examination : 05/10/2025 (FN)
 Exam Duration : 90 minutes Maximum Marks: 50

General instruction(s):

- Answer All Questions
- M - Max mark; CO - Course Outcome; BL - Blooms Taxonomy Level (1 - Remember, 2 - Understand, 3 - Apply, 4 - Analyse, 5 - Evaluate, 6 - Create)
- 1. Course Outcomes:
 2. Explain the detailed architecture, define objects, load data, query data and performance tune Document-oriented NoSQL databases.
 3. Demonstrate an understanding of the detailed architecture, define objects, load data, query data and performance tune column oriented and Graph oriented NoSQL Databases

Q.No	Question	Module	Marks	CO	BL
1	a) A ride-sharing company operates in multiple cities across India and stores massive amounts of trip data in a document-oriented database. Each trip document contains the following fields: • trip_id (unique identifier for each ride) • driver_id (ID of the driver assigned) • rider_id (ID of the customer) • city (location of the ride, e.g., Chennai, Mumbai, Delhi) • fare_amount (total cost of the trip) • trip_date (date and time of the ride) As the system scales, query performance and efficient storage become critical. Different stakeholders require data access in different ways: 1. Range Partitioning Requirement: Managers frequently analyze time-based reports such as total trips in the last week, last month, or quarterly performance. 2. Hash Partitioning Requirement: Since millions of trips are generated daily, there is a risk of data skew if trips are not evenly distributed across storage nodes. 3. List Partitioning Requirement: Regional managers access trips primarily based on city, e.g., Chennai, Mumbai, Delhi. Demonstrate how each partitioning can be applied to the dataset and explain which attribute can be used as partition key?	4	5	3	2
	b) A large e-commerce company stores millions of orders in its database. A manager wants to run the following query: <pre>SELECT order_id, customer_name, product_type, order_date FROM Orders WHERE order_date >= '2025-01-01' AND product_type = 'Electronics';</pre>	4	5	3	2



SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2025-2026

SLOT: A2-1A2

	The Orders table has: • 50 million records in total • 40% of records are from 2025 onwards • Only 5% of all orders are of product_type = "Electronics" 1. How a query processor can be used to optimize the execution of this query? 2. Which condition should be applied first, and why? 3. How does this minimize the data scanned and improve query performance? 4. What role do indexes play in further optimization?			
2	A retail company wants to run a marketing campaign targeting customers who have purchased home electronics in the past 12 months. The company's database contains the following tables: • Customer: customer_id, name, email, address • Orders: order_id, customer_id, order_date • Order_Items: order_item_id, order_id, product_id, quantity • Products: product_id, product_name, product_category, price The marketing team wants to generate a report listing: • Customer name and email • Products purchased in the electronics category • Quantity purchased • Order date Tasks: 1. Explain how you would use joins to retrieve this data efficiently. 2. Identify which attributes serve as foreign keys and how they connect the tables. 3. Describe potential performance issues when joining multiple large tables and suggest optimization techniques that the database query processor might use. 4. Draft a sample NoSQL query that retrieves the required information.	4	10	3
3	A healthcare analytics company, MediTrack, is building a patient monitoring system on top of Google Bigtable. The system stores real-time patient vital signs such as heart rate, blood pressure, and temperature from wearable IoT devices. The solution must handle billions of data points every day and provide near real-time insights to doctors. As the lead data engineer for MediTrack: 1. How would you design the row key to meet the requirements of efficient indexing and fast range queries? 2. Which core feature of Bigtable ensures that adding a new vital sign column (like oxygen saturation) does not require schema changes? 3. How can you control data location so that patients from the same hospital region have their data stored closer together? 4. What mechanism in Bigtable guarantees that an update to a patient's record does not result in partial or inconsistent data? 5. Why is the sorted row feature of Bigtable crucial for running queries over a patient's vitals within a time range?	5	10	



SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2025-2026

4	A global e-commerce company, ShopGlobal, is planning to build a large-scale recommendation and analytics system. The system must handle: • Millions of concurrent users browsing product catalogs in real time. • Write-heavy workloads from order transactions, clickstreams, and user behavior logs. • Global deployment across multiple data centers to ensure low latency for users worldwide. • High availability — the system must keep running even if some nodes or data centers fail. • Scalable architecture that can grow seamlessly as data volume increases. The engineering team is evaluating Apache Cassandra and Apache HBase as possible database solutions. As the lead architect, justify why Cassandra's architecture is more advantageous than HBase's architecture for ShopGlobal's system. Explain how the following enhance Cassandra's availability and performance compared to HBase: • Commit Log (for durability and crash recovery) • Bloom Filter (for efficient data lookup without unnecessary disk reads) • Consistency Level (for tunable trade-off between consistency, availability, and performance)	5	10	4	3
5	Design a Cassandra database with CQL queries to store student course enrollments with fields: student_id, student_name, course_id, course_name, semester, grade. a) Design an enrollments table so that queries like "all courses taken by a student ordered by semester" are efficient. Justify your choice of partition key and clustering key. b) Insert a record for student S2101, name Anita Sharma, enrolled in course CSE501, name Distributed Systems, semester Spring2024, grade A. c) Since Cassandra does not support joins, propose how you would design the schema or queries to efficiently retrieve "all students who took Distributed Systems in Spring2024". d) Delete all enrollments of student S2101 before Spring2022, considering the clustering key. e) Write a query to fetch the latest course enrollment (by semester) for student S2101.	6	10	4	3
