

**VIT**Vellore Institute of Technology
(Approved by the Government of Tamil Nadu under Section 3, FET Act, Act 10 of 1964)**Slot : F2+TF2****School of Computer Science Engineering and Information Systems****Winter Semester 2024-2025****Continuous Assessment Test – II****Programme Name & Branch: B.Tech., Information Technology****Course Name & code: BITE411L Big Data Analytics****Class Number (s): VL2024250502672, VL2024250502679 & VL2024250502647****Faculty Name (s): Dr.Pounambal, Dr.Balasubramani & Dr.Mohanapriya****Exam Duration: 90 Min.****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)
- Course Outcomes:
CO4: Process Data using Spark and No SQL Databases
CO5: Apply MapReduce based analysis.

Q.No	Question	Max Marks	CO	BL																														
1.	Amazon has one petabyte of customer wish list for over three years. Every thirty minutes new wish list gets added to the database. Illustrate how Lambda architecture can be developed for the above requirement to make suitable recommendations for its customer.	10	CO 4	BL 3																														
2.	Consider we want to count the number of words in text data received from a data server listening on a TCP socket. Construct the Spark workflow with the driver program using RDD and explain the components involved.	10	CO 4	BL 3																														
3.	Write the CQL query for the following. a) Create the table in Cassandra which will store the following sensor data.	10	CO 4	BL 3																														
<table border="1"> <thead> <tr> <th>Machine_id</th><th>Sensor_id</th><th>Event_timestamp</th><th>Observation type</th><th>Observation value</th></tr> </thead> <tbody> <tr> <td>c11e4923</td><td>4704g26</td><td>2025-01-10 10:00:00</td><td>Temperature</td><td>72.5</td></tr> <tr> <td>1b1e4871</td><td>222e803</td><td>2025-03-10 10:01:00</td><td>Pressure</td><td>1012.3</td></tr> <tr> <td>111c4567</td><td>333e467</td><td>2025-03-10 10:02:00</td><td>Humidity</td><td>45.2</td></tr> <tr> <td>444e4582</td><td>235e421</td><td>2025-03-10 10:05:00</td><td>Temperature</td><td>68.9</td></tr> <tr> <td>a4564265</td><td>55e6517</td><td>2025-03-10 10:06:00</td><td>Pressure</td><td>1010.5</td></tr> </tbody> </table>					Machine_id	Sensor_id	Event_timestamp	Observation type	Observation value	c11e4923	4704g26	2025-01-10 10:00:00	Temperature	72.5	1b1e4871	222e803	2025-03-10 10:01:00	Pressure	1012.3	111c4567	333e467	2025-03-10 10:02:00	Humidity	45.2	444e4582	235e421	2025-03-10 10:05:00	Temperature	68.9	a4564265	55e6517	2025-03-10 10:06:00	Pressure	1010.5
Machine_id	Sensor_id	Event_timestamp	Observation type	Observation value																														
c11e4923	4704g26	2025-01-10 10:00:00	Temperature	72.5																														
1b1e4871	222e803	2025-03-10 10:01:00	Pressure	1012.3																														
111c4567	333e467	2025-03-10 10:02:00	Humidity	45.2																														
444e4582	235e421	2025-03-10 10:05:00	Temperature	68.9																														
a4564265	55e6517	2025-03-10 10:06:00	Pressure	1010.5																														

	b) Load the data in to the table. c) Retrieve data for specific time range. d) Update the observation value for specific sensor e) Delete the data for specific timestamp.																																																																														
4.	Evaluate Parellel Apriori algorithm in detail for the following super market scenario. Let support count as 2 and confidence as 60%. <table><tr><td>Transacti on ID</td><td>Onion</td><td>Potato</td><td>Burger</td><td>Milk</td><td>Gee</td></tr><tr><td>T1</td><td>1</td><td>1</td><td>1</td><td>0</td><td>0</td></tr><tr><td>T2</td><td>0</td><td>1</td><td>1</td><td>1</td><td>0</td></tr><tr><td>T3</td><td>0</td><td>0</td><td>0</td><td>1</td><td>1</td></tr><tr><td>T4</td><td>1</td><td>1</td><td>0</td><td>1</td><td>0</td></tr><tr><td>T5</td><td>1</td><td>1</td><td>1</td><td>0</td><td>1</td></tr><tr><td>T6</td><td>1</td><td>1</td><td>1</td><td>1</td><td>1</td></tr></table>	Transacti on ID	Onion	Potato	Burger	Milk	Gee	T1	1	1	1	0	0	T2	0	1	1	1	0	T3	0	0	0	1	1	T4	1	1	0	1	0	T5	1	1	1	0	1	T6	1	1	1	1	1	10	CO 5	BL 4																																	
Transacti on ID	Onion	Potato	Burger	Milk	Gee																																																																										
T1	1	1	1	0	0																																																																										
T2	0	1	1	1	0																																																																										
T3	0	0	0	1	1																																																																										
T4	1	1	0	1	0																																																																										
T5	1	1	1	0	1																																																																										
T6	1	1	1	1	1																																																																										
5.	The following table consists of training data. Find the best feature at the root and draw the final decision tree. Use impurity measures of Entropy and Information gain to find the best feature. <table><tr><td>ID</td><td>Fever</td><td>Cough</td><td>Breathing issues</td><td>Infected (Target)</td></tr><tr><td>1</td><td>No</td><td>No</td><td>No</td><td>No</td></tr><tr><td>2</td><td>Yes</td><td>Yes</td><td>Yes</td><td>Yes</td></tr><tr><td>3</td><td>Yes</td><td>Yes</td><td>No</td><td>No</td></tr><tr><td>4</td><td>Yes</td><td>No</td><td>Yes</td><td>Yes</td></tr><tr><td>5</td><td>Yes</td><td>Yes</td><td>Yes</td><td>Yes</td></tr><tr><td>6</td><td>No</td><td>Yes</td><td>No</td><td>No</td></tr><tr><td>7</td><td>Yes</td><td>No</td><td>Yes</td><td>Yes</td></tr><tr><td>8</td><td>Yes</td><td>No</td><td>Yes</td><td>Yes</td></tr><tr><td>9</td><td>No</td><td>Yes</td><td>Yes</td><td>Yes</td></tr><tr><td>10</td><td>Yes</td><td>Yes</td><td>No</td><td>Yes</td></tr><tr><td>11</td><td>No</td><td>Yes</td><td>No</td><td>No</td></tr><tr><td>12</td><td>No</td><td>Yes</td><td>Yes</td><td>Yes</td></tr><tr><td>13</td><td>No</td><td>Yes</td><td>Yes</td><td>No</td></tr><tr><td>14</td><td>Yes</td><td>Yes</td><td>No</td><td>No</td></tr></table>	ID	Fever	Cough	Breathing issues	Infected (Target)	1	No	No	No	No	2	Yes	Yes	Yes	Yes	3	Yes	Yes	No	No	4	Yes	No	Yes	Yes	5	Yes	Yes	Yes	Yes	6	No	Yes	No	No	7	Yes	No	Yes	Yes	8	Yes	No	Yes	Yes	9	No	Yes	Yes	Yes	10	Yes	Yes	No	Yes	11	No	Yes	No	No	12	No	Yes	Yes	Yes	13	No	Yes	Yes	No	14	Yes	Yes	No	No	10	CO 5	BL 4
ID	Fever	Cough	Breathing issues	Infected (Target)																																																																											
1	No	No	No	No																																																																											
2	Yes	Yes	Yes	Yes																																																																											
3	Yes	Yes	No	No																																																																											
4	Yes	No	Yes	Yes																																																																											
5	Yes	Yes	Yes	Yes																																																																											
6	No	Yes	No	No																																																																											
7	Yes	No	Yes	Yes																																																																											
8	Yes	No	Yes	Yes																																																																											
9	No	Yes	Yes	Yes																																																																											
10	Yes	Yes	No	Yes																																																																											
11	No	Yes	No	No																																																																											
12	No	Yes	Yes	Yes																																																																											
13	No	Yes	Yes	No																																																																											
14	Yes	Yes	No	No																																																																											