VIT

Slot : G2+TG2

School of Computer Science Engineering and Information Systems

FALL Semester 2025-2026

Continuous Assessment Test – II

Programme Name & Branch: B.Tech. IT **Class Number (s):** VL2025260107397

Faculty Name (s): Dr. Swarna Priya RM **Course Name & code:** BITE411L BIG DATA ANALYTICS

Exam Duration: 90 Min. **Maximum Marks:** 50

Q.No.	Question	Max Marks	CO	BL
1 (a)	A media company processes video logs (huge files) daily. With classic MapReduce, jobs frequently fail due to imbalanced load distribution. How would YARN's ApplicationMaster and Container model improve data locality and job performance in this scenario? Compare how classic MapReduce vs. YARN allocates resources in this workload. Which system is more efficient and why?	5	CO5	BL4
(b)	Your organization runs predictive healthcare analytics on a 50-node Hadoop cluster using classic MapReduce. As the workload grows to 500 nodes, frequent job delays and failures occur. Identify the scalability bottlenecks in classic MapReduce and justify how YARN improves performance.	5		
2	Your company has been running batch-only Hadoop MapReduce jobs for the past 5 years to process daily sales and customer analytics. Due to increasing demand for real-time insights, management has decided to migrate to Apache Spark on YARN, enabling both batch and streaming workloads. As the lead data engineer, you are asked to evaluate the migration challenges. Discuss the key challenges you would face in the following aspects, and suggest strategies to overcome them. • How will YARN's Capacity Scheduler/Fair Scheduler differ from Hadoop's FIFO job scheduling? What issues may arise when real-time streaming jobs and batch jobs compete for cluster resources? • How can dynamic allocation and container reuse be used to mitigate scheduling conflicts?	10	CO4	BL4
3	A university wants to predict whether a student will Pass (P) or Fail (F) a course based on two attributes, Study Hours (H): {Low, High} Attendance (A): {Poor, Good} The training dataset of 10 students is shown below: (i) Compute the entropy of the target class (ii) Compute the information gain of Study Hours (H) and Attendance (A). (iii) Identify which attribute should be used as the root node. (iv) Draw the decision tree up to one level (root + child splits). (v) Use the decision tree to predict the class for a new student S11 with: Study Hours = Low, Attendance = Good	10	CO3	BL3

Student	Study Hours (H)	Attendance (A)	Result
S1	Low	Poor	F
S2	Low	Good	F
S3	High	Poor	P
S4	High	Good	P
S5	Low	Poor	F
S6	High	Good	P
S7	High	Poor	P
S8	Low	Good	F
S9	High	Good	P
S10	Low	Poor	F

4.	You are analyzing sales transactions for a small e-commerce platform. The transactions are stored across 3 servers (nodes). Each transaction records the items purchased by a customer in one order. **Node 1 Transactions:** 1. {Laptop, Mouse, Keyboard} 2. {Laptop, Mouse} 3. {Laptop, Keyboard} 4. {Mouse, Keyboard} 5. {Laptop, Mouse, Keyboard} **Node 2 Transactions:** 1. {Laptop, Keyboard} 2. {Mouse, Keyboard} 3. {Laptop, Mouse} 4. {Laptop, Keyboard} 5. {Mouse, Keyboard} **Node 3 Transactions:** 1. {Laptop, Mouse, Keyboard} 2. {Mouse, Keyboard} 3. {Laptop, Keyboard} 4. {Laptop, Mouse} 5. {Keyboard} **Global Minimum Support:** 6 **Tasks:** 1. **Phase 1 (Local Frequent Itemsets)**: ○ Compute **1-itemset** and **2-itemset** supports locally for each node. ○ Identify **locally frequent itemsets** (support ≥ local threshold, which can be proportional to node size). 2. **Phase 2 (Global Frequent Itemsets)**: ○ Merge local frequent itemsets to generate **candidate global itemsets**. ○ Count **global support** for each candidate. ○ Identify **globally frequent itemsets**. 3. Analysis: ○ Which items are frequently bought together? ○ If the platform wants to bundle products, which **2-item combinations** are the best candidates?	10	CO3	BL3
5.	You have a dataset of 6 samples. | Sample	Feature X	Feature Y	Class
---	---	---	---
1	2	3	A
2	3	4	A
3	4	3	B
4	5	6	B
5	6	5	B
6	7	6	A | You train a **Random Forest** with **3 decision trees**, using bootstrap samples and max 2 features per split. **Tasks:** 1. Generate **3 bootstrap samples** (you can select any random subsets with replacement). 2. Draw simple **decision trees** for each bootstrap sample (1-2 splits only). 3. Predict the class for a **new sample (4,5)** using **majority voting**.	10	CO3	BL3