

JK Lakshmipat University Jaipur
INSTITUTE OF ENGINEERING AND TECHNOLOGY
 End Term Examination, December 2025
 B.Tech/BCA, Semester V, 2025-26*

Roll No.....

CS1243: Big Data Analytics with Hadoop and Spark

Time: 2.5 hours

Max. Marks: 30

Instructions to students:

1. Do not write anything other than your roll number on question paper.
2. Mention all the assumptions for your answers clearly.
3. **Scientific Calculators are permitted**

Q.1	Consider the following scenarios. For each, give your ideal solution and explain the idea behind arriving at the solution. 1. You are running an RDD join operation between a massive RDD (Orders) and a small reference RDD (Products). The job is hanging, and the Spark UI shows one or two tasks taking far too long. What is the likely issue, and what RDD operation could resolve this? 2. You have a long sequence of transformations on an RDD: $R_0 \rightarrow R_1 \rightarrow R_2 \rightarrow R_3$. If you call <code>R_3.count()</code>, what happens to the computation for R_1 and R_2? 3. You notice that a <code>groupByKey()</code> operation is causing excessive memory pressure and spilling to disk on a few executors. How can you rewrite this using an RDD transformation to improve performance? 4. If you use <code>.collect()</code> on a massive RDD, what is the immediate risk to the Driver Program? 5. You read a large table into a DataFrame, perform a non-key-based filter (<code>.filter(col("status") == 'A')</code>), and then immediately save the result to S3. You notice an unnecessary shuffle occurring. <i>Why does it occur? How do you fix it?</i>	2 2 1 1 2 (CO6)
Q1.2	A Spark cluster contains one driver and three executor nodes. A dataset of 1 million two-dimensional points is stored in an RDD/DataFrame and is evenly partitioned across the three executors. The K-Means algorithm (with $k = 2$) is run on this cluster. Provide a step-by-step distributed execution trace for one iteration of K-Means on Spark. Your diagrams and explanation should show the flow of data and control clearly marking the map phase and reduce phase.	4 (CO5)

Q.2	(a) Consider the following social network graph $G = (V,E)$, $V = \{1,2,3,4,5,6\}$ and bi-directional edges $E=\{(1,2), (1,3), (2,4), (4,6), (3,5), (3,4), (5,6)\}$. Starting with node 1, perform the first iteration of the Girvan-Newman algorithm and arrive at the edge flows. OR Consider the following social network graph $G = (V,E)$, $V = \{1,2,3,4\}$ and bi-directional edges $E=\{(1,2), (1,3), (1,4), (2,3), (3,4)\}$. Perform the first iteration of the NOVER algorithm and calculate the NOVER scores. What are the communities formed?	2												
	(b) The modularity of a community with a single node is _____.	1												
	(c) A Spark cluster contains one driver and two executor nodes. A graph with edges $A \rightarrow B$, $B \rightarrow C$ and $A \rightarrow D$ is stored in an RDD/DataFrame. Assume a node-update rule of your choice. Provide a step-by-step distributed execution of this graph in the spark cluster.	3 (CO7)												
Q.3	(a) Consider the following data stream <table border="1"><tr><td>Time</td><td>10:55:00</td><td>10:55:06</td><td>10:55:07</td><td>10:55:15</td><td>10:55:16</td></tr><tr><td>Event</td><td>23</td><td>10</td><td>5</td><td>77</td><td>33</td></tr></table> 1. Find the MAX value every 4 seconds using Tumbling window 2. Find the MAX value for a window size of 4 secs with a hop size of 3 secs 3. Find the MAX value for a sliding window of 9 secs	Time	10:55:00	10:55:06	10:55:07	10:55:15	10:55:16	Event	23	10	5	77	33	1 1 1
Time	10:55:00	10:55:06	10:55:07	10:55:15	10:55:16									
Event	23	10	5	77	33									
	(b) A Spark Streaming job (batch interval = 10 seconds) monitors a directory, named <i>input</i>, for new text files. After the job begins, a file named file1.txt is added to the directory. Write a Spark Streaming program that counts the number of lines received in the last 30 seconds, updating the count every 10 seconds. (Hint: It is a stateful operation)	3 (CO7)												

Q.4	Consider the following log file, log.csv: <pre>UserID,ActionType,DurationSeconds,Timestamp U031,view,56,2024-01-01 00:00:06 U048,login,0,2024-01-01 00:00:09 U031,login,0,2024-01-01 00:00:18 U007,purchase,1,2024-01-01 00:00:21 U037,logout,0,2024-01-01 00:00:27 U049,login,0,2024-01-01 00:00:29 U047,view,92,2024-01-01 00:00:31</pre> Assume your data is in rdd, named logRdd or in a data frame name logDF. Write transformations and actions for the following: 1. Using rdd: Determine the total number of 'purchase' ActionType recorded in the dataset. 1 2. Using rdd: Calculate the total time (in seconds) each unique user spent performing activities 1 3. Using rdd: Find the Top 3 UserIDs with the highest total activity duration. 2 4. Using data frames: Calculate the average duration (in seconds) of the 'view' ActionType per userID 1 5. Using data frames: Find the Top 3 UserIDs with the highest total activity duration. 1	(CO6)
-----	--	-------

