BIRLA INSTITUTE OF TECHNOLOGY, MESRA, RANCHI
(END SEMESTER EXAMINATION)

CLASS: IMSC **SEMESTER : IX**
BRANCH: QEDS **SESSION : MO/2025**

SUBJECT: ED547 BIG DATA ANALYTICS

TIME: 3 Hours **FULL MARKS: 50**

INSTRUCTIONS:
1. The question paper contains 5 questions each of 10 marks and total 50 marks.
2. Attempt all questions.
3. The missing data, if any, may be assumed suitably.
***4*. Before attempting the question paper, be sure that you have got the correct question paper.**
5. Tables/Data hand book/Graph paper etc. to be supplied to the candidates in the examination hall.

			CO	BL
Q.1(a)	Explain why data veracity and data value are critical in Big Data analytics	[2]	2	3
Q.1(b)	Identify and explain at least five key challenges in Big Data management	[3]	2	2
Q.1(c)	An online shopping platform wants to analyse customer browsing and purchase history to recommend products. • Describe the Big Data pipeline for this system from data collection to recommendation generation. • Which Big Data challenges must be addressed to ensure accurate and timely recommendations?	[5]	4	3
Q.2(a)	Assume a 128 MB block size and replication factor of 3. If a cluster has 6 nodes and a 1 GB file is stored, explain precisely how data blocks are distributed across the nodes.	[2]	2	3
Q.2(b)	How does HBase's column-oriented storage model differ from traditional row-oriented databases?	[3]	1	2
Q.2(c)	Why is HDFS not ideal for low-latency data access? Discuss in terms of HDFS architecture, write pipeline, and block placement strategy.	[5]	1	2
Q.3(a)	Elaborate the main features of RDD in Apache Spark and explain each of them briefly with an example.	[2]	2	2
Q.3(b)	Consider an RDD input_rdd with the following elements: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] Perform the following operations on the RDD and obtain the output: • Filter out the even numbers. • Square the remaining numbers. • Compute the sum of squares. Write the code to perform these operations on the RDD and obtain the output.	[3]	3	3/4
Q.3(c)	Consider a file (dataset) containing numeric values, write steps to perform the MapReduce job and formulate the MapReduce job to calculate the mean? Consider the file (dataset) as given below: 96,54 12,60 29,52	[5]	4	3/4
Q.4(a)	Differentiate between Sqoop and Flume	[2]	1	2
Q.4(b)	Explain why data preprocessing is critical before Big Data analytics. What are the main challenges in cleaning unstructured or semi-structured data?	[3]	3	2
Q.4(c)	Define ETL (Extract, Transform, Load). How does ETL differ in traditional systems versus Big Data pipelines (e.g., Spark or Hadoop-based ETL)?	[5]	3	2
Q.5(a)	Explain the Pig architecture in detail and how it differs from traditional MapReduce programming.	[5]	4	3
Q.5(b)	Explain the various components of the Hive architecture and how they work together to execute queries.	[5]	4	2

:::::26/11/2025:::::E