

COURSE CODE (CREDITS): 25B1WCI741 (3)

MAX. MARKS: 15

COURSE NAME: Big Data & Real-Time Analytics

MAX. TIME: 60 Min.

COURSE INSTRUCTORS: D. Gupta

Note: 1) All questions are compulsory. Marks and COs for each question are indicated. 2) Answer the questions in the given order. 3) Be concise and write neatly. 4) Use of calculators is not allowed.

Q. No.	Question	CO	Marks
Q. 1	Explain the 5 V's of Big Data (Volume, Velocity, Variety, Veracity, Value) with one real-world example each. Why is Veracity the most commonly underestimated V, and how does poor veracity undermine value even when volume and velocity are high?	1	3
Q. 2	Explain the curse of dimensionality. Why increasing the number of features (dimensions) in a dataset causes distances between data points to become less meaningful, and why this breaks distance-based methods like k-NN and clustering. Using the concept of sparsity and the empty space phenomenon, briefly explain why a dataset with thousands of features needs far more samples to remain statistically reliable than a dataset with only a few features.	2	3
Q. 3	A retail company wants to build a pipeline that (i) stores years of historical transaction logs for periodic fraud-model retraining, and (ii) scores incoming transactions for fraud in real time, within milliseconds. Propose a suitable architecture for this system, choosing between HDFS vs. cloud object storage for the historical layer, and batch (MapReduce/Apache Spark) vs. streaming (Apache Kafka/Apache Spark Streaming) for the real-time layer. Justify each choice and briefly explain how the batch and streaming components would work together.	2	3
Q. 4	a) Explain the roles of the NameNode and DataNode in HDFS architecture. Why is losing the NameNode's metadata more catastrophic than losing a single DataNode, even though the DataNode holds the actual data? b) A company ingests 4 TB of log data per day using HDFS's default block size 128 MB and default replication factor 3. Calculate (i) the number of blocks created per day, and (ii) the raw cluster storage consumed per day. c) Explain how HDFS achieves fault tolerance when a DataNode fails. Describe the sequence from heartbeat loss to re-replication, and state why clients typically don't notice the failure.	1	6 (2+2+2)