SNS COLLEGE OF TECHNOLOGY

(An Autonomous Institution)

23MCT305 – DATA ANALYTICS IN AUTOMATION SYSTEM

TECHNICAL PUZZLES

UNIT IV: MANAGING HADOOP FILE (CO4)

Puzzle No.	Puzzle Description	Expected Approach / Key Concept	Bloom's Level	CO
1	A 300 GB file is stored in HDFS with block size 128 MB and replication factor 3. Calculate total storage used across the cluster and explain what happens if one DataNode fails.	300/128 ≈ 3 blocks → 9 copies → 900 GB; data still accessible	Apply	CO4
2	You have a choice: upgrade 10 nodes with better CPU/RAM (vertical) or add 20 new average nodes (horizontal). For a MapReduce job that is I/O bound, which would improve performance more and why?	Horizontal (more parallelism in Map phase)	Analyze	CO4
3	In Pseudo Distributed mode, all daemons run on one machine. Argue why this mode is preferred for learning MapReduce programming over Stand-alone mode, despite similar hardware use.	Simulates real cluster behavior (NameNode, DataNode separate processes)	Evaluate	CO4
4	A MapReduce job fails because NameNode runs out of memory. Suggest two architectural solutions using only Hadoop concepts from Unit IV.	Add Secondary NameNode, increase replication cautiously	Analyze	CO4
5	Design a simple fault-tolerant strategy: if a TaskTracker fails mid-job, how does Hadoop recover? Describe the sequence of events involving JobTracker.	JobTracker reassigns task to another node	Create	CO4