ourdiu.com | Shared by MD. GOLAM MUSTAKIM NAEEM

Daffodil International University (DIU)

Department of Computing and Information System (CIS)

Course: Big Data Analytics and Applications (CIS312) **Exam:** Class Test 03 (Set A)
Time: 30 Minutes **Total Marks:** 15

Answer all questions. Be concise and provide architectural justifications where required.

1. **[Marks: 3 + 2 = 5]**
   a) In legacy Hadoop (MRv1), a node configured with separate Map and Reduce slots often faced severe resource underutilization. Explain how YARN resolves this issue by introducing dynamic **Containers**.
   b) Hadoop MapReduce writes intermediate outputs to the disk after every stage. Briefly explain why this disk I/O bottleneck makes iterative machine learning models (like K-Means) extremely slow compared to Apache Spark.

2. **[Marks: 2 + 3 = 5]**
   a) To ensure fault tolerance, HDFS stores 3 physical copies (replication) of every data block. Explain how Spark's **RDD Lineage** achieves fault tolerance without relying on multiple physical copies in the RAM.
   b) In Spark DAG execution, distinguish between **Narrow Dependency** and **Wide Dependency**. Which of these dependencies triggers a network shuffle and creates a new Stage boundary?

3. **[Marks: 2 + 3 = 5]**
   a) MapReduce enforces a rigid two-step execution model (Map followed by Reduce). Briefly explain how Apache Spark's **Directed Acyclic Graph (DAG)** approach provides more flexibility for complex, multi-step data processing pipelines.
   b) Why is executing the collect() Action on a multi-terabyte RDD considered a critical anti-pattern? Describe its direct impact on the Spark **Driver's** memory.

Why is executing the collect() on a multiterabyte RDD considered critical anti-pattern –