WEEK 14 QUIZ SOLUTION

1. Which of these is true about number of partitions generated post shuffling, for structured API in Spark?
   a. ***200 partitions by default - can be changed**
   b. 200 partitions by default - can't be changed
   c. Depends on I/P file size and hdfs block size
   d. None

   Explanation: Structured API uses default value of spark.sql.shuffle.partitions is 200 , which can be changed as per the requirement.

2. Suppose we have a small dataset and a very large dataset, which of these when used can provide significant optimization in spark? (Multiple options can be true.)
   a. Broadcast the larger dataset across all executors
   b. ***Broadcast the smaller dataset across all executors**
   c. Infer schema explicitly for smaller dataset
   d. ***Infer schema explicitly for larger dataset**

   Explanation: B and D are true.

   - Broadcast is done using small table because its copied across executors.
   - Infer schema takes time for large dataset hence its recommended to explicitly mention the schema

3. For which of these, when used can reduce the partition skewing?
   a. ***Repartition**
   b. Coalesce

   Explanation: Repartition can solve the skew problem because the result partition will be of similar size.

4. To reduce the number of partitions which is favored and why?
   a. Repartition, avoids full shuffling
   b. ***Coalesce, avoids full shuffling**

c. Repartition, avoids local shuffling
d. Coalesce, avoids local shuffling

Explanation: Coalesce because it does local shuffle only.

5. While using spark-submit command, the default deploy mode, when otherwise not specified exclusively is _______
   a. *Client Mode
   b. Cluster Mode
   c. No default
   d. deploy-mode should be exclusively specified

   Explanation: Default deploy mode is client

6. When we run spark submit in cluster mode, the results for the collect action can be viewed in
   a. gateway node terminal
   b. *worker node standard output logs

   Explanation: In cluster mode , driver runs on one of the worker node where you can see the logs.

7. Which of these is not a proper optimization technique? (Multiple can be chosen)
   a. Increase Cardinality to Maximize Parallelism
   b. *Do filtering post shuffle phase
   c. Decrease number of Skewed-Partitions
   d. *Join two large datasets using Broadcast join

   Explanation: B and D are not a proper optimization technique.

   - Filtering must be done before shuffle , this way we minimize the data send to shuffle.
   - Broadcast join needs one small table which gets replicated on every executor machine.

8. We want to join two large dataframes. Consider spark.sql.shuffle.partitions default value. We have 30 executors with 8 CPU cores each and number of distinct keys in join column is 220. So what will be the degree of parallelism at max at this point?
    a. 220
    b. 240
    c. 30
    d. *None of the options

    Explanation: min(Total Cores, Number of Shuffle Partitions, Number of Distinct Keys)

    = min(240, 200, 220)

    = 200

9. Which of these is TRUE about hash aggregate? (Multiple can be chosen)
    a. *Skips sorting of data internally
    b. Time complexity is O(nlogn) ,n being number of records
    c. *Extra space required, depends on number of distinct keys
    d. Memory for hash table is part of the Executor Memory
    e. All columns datatypes in value for a key-value pair, should be immutable.

    Explanation: A and C are true

10. By default spark always internally tries to apply hash aggregate whenever possible.
    a. *TRUE
    b. FALSE

11. User can add their own rules in Catalyst Optimizer.
    a. *TRUE
    b. FALSE

12. Table name mismatch will be caught in which of these stages?
    a. Unresolved Logical Plan
    b. Optimized Logical Plan
    c. *Analyzed Logical Plan
    d. Physical Plan

13. Which aggregate to be used internally is decided in
    a. Unresolved Logical Plan
    b. Optimized Logical Plan
    c. Analyzed Logical Plan
    d. *Physical Plan

14. Consider you have 2 large files using dataframes, spark.sql.shuffle.partitions default value. We have 20 executors with 4 CPU cores each and number of distinct keys in join column is 40. So what will be the degree of parallelism at max at this point?
    a. 80
    b. 200
    c. *40
    d. 160

    Explanation: min(Total Cores, Number of Shuffle Partitions, Number of Distinct Keys)

    =min(80,200,40)

    = 40

15. Where is the driver program running : spark2-submit \ --class LogLevelGrouping \ --master yarn \ --deploy-mode cluster \ --executor-memory 3G \ --num-executors 4 \wordcount.jar bigLogNew.txt
    a. Client node
    b. *Worker node
    c. Edge node
    d. Data node

Explanation: In cluster mode , driver runs on worker node.

16. Dataframe can connect to external datasource like mysql

    a. *TRUE
    b. FALSE

17. Sort Aggregate is faster than Hash Aggregate

    a. TRUE
    b. *FALSE

18. Catalyst optimizer will optimize the execution plan for RDD

    a. TRUE
    b. *FALSE

19. Syntax errors are checked in ______ plan

    a. *Parsed Logical
    b. Analyzed logical
    c. Physical
    d. None of the above

20. When we are using hash aggregate we should have mutable types in the values

    a. *TRUE
    b. FALSE