DS 644: Homework 4

Instructions. Answer the following multiple choice questions by selecting all correct choices.

1. (4 points) **Shuffling**.

   (a) What is *shuffling*?

   - ☐ a method for recovering data after hardware failure
   - ☐ the method used to ensure a random number generator is unbiased
   - ☐ any movement of data
   - ☐ moving data from memory to disk, usually caused by insufficient fast memory
   - ☐ transferring data between nodes in a cluster, usually in order to complete a computation

   (b) How can shuffling sometimes be reduced or avoided using Spark?

   - ☐ use higher quality, fault-tolerant hardware
   - ☐ use a pre-shuffled random number generator
   - ☐ avoid algorithms that process the entire data set in favor of algorithms that only need a small subset of it
   - ☐ use only fast memory, eliminating all spinning disks from the network
   - ☐ partition an `RDD` before applying transformations or actions that cause shuffling

2. **Partitions and Partitioning**

   (a) Given a pair RDD (of key-value pairs), when we group values with the same key Spark collects key-value pairs with the same key on the same machine of our cluster.

   ☐ True ☐ False

   (b) By default, Spark uses range partitioning to determine which key-value pair should be sent to which machine.

   ☐ True ☐ False

   (c) Suppose we partition an RDD into a number of blocks. From the following statements, select the two that are true.

   - ☐ A single block of the partition may be distributed across multiple machines in the cluster.
   - ☐ A block of the partition is assigned to at most one machine of the cluster.
   - ☐ At least one block of the partition is assigned to every machine in the cluster.
   - ☐ At most one block of the partition is assigned to every machine in the cluster.
   - ☐ More than one block of the partition may be assigned to the same machine in the cluster.

3. Consider a Pair RDD, with keys [8, 23, 39, 40, 97], and suppose we want to partition these data into 4 blocks.

   Using hash partitioning with the identity as `hashCode()` function (`n.hashCode() == n`), check the boxes next to the numbers assigned to the given partition block.

   (a) block 0: ☐ 8 ☐ 23 ☐ 39 ☐ 40 ☐ 97 ☐ none

   (b) block 1: ☐ 8 ☐ 23 ☐ 39 ☐ 40 ☐ 97 ☐ none

   (c) block 2: ☐ 8 ☐ 23 ☐ 39 ☐ 40 ☐ 97 ☐ none

   (d) block 3: ☐ 8 ☐ 23 ☐ 39 ☐ 40 ☐ 97 ☐ none

4. Consider a Pair RDD, with keys [8, 23, 39, 40, 97], and suppose we want to partition these data into 4 blocks.

   Using range partitioning with ranges [0, 20], [21, 40], [41, 60], [61, 100], check the boxes next to the numbers assigned to the given partition block.

   (a) block 0: ☐ 8 ☐ 23 ☐ 39 ☐ 40 ☐ 97 ☐ none

   (b) block 1: ☐ 8 ☐ 23 ☐ 39 ☐ 40 ☐ 97 ☐ none

   (c) block 2: ☐ 8 ☐ 23 ☐ 39 ☐ 40 ☐ 97 ☐ none

   (d) block 3: ☐ 8 ☐ 23 ☐ 39 ☐ 40 ☐ 97 ☐ none

5. Which strategy would result in a more balanced distribution of the data across the partition?

   ☐ hash partitioning ☐ range partitioning

6. (a) Which method can we use to determine whether Spark recognizes that a transformation or action will result in shuffling?

☐ `debugDAG` ☐ `isShuffled` ☐ `showSchema` ☐ `showExecutionPlan` ☐ `toDebugString`

(b) How data is initially partitioned and arranged on the cluster doesn't matter, since Spark will always re-arrange your data to avoid shuffling.

☐ True ☐ False

(c) `reduceByKey` running on a pre-partitioned ROD will computed values locally, requiring only the final reduced values to be sent from workers to the driver.

☐ True ☐ False

(d) `join` called on two RDDs that are pre-partitioned with the same partitioner and cached on the same node will cause the join to be computed locally, with no shuffling across the network.

☐ True ☐ False

(e) Suppose algorithm **A** joins two RDDs and then performs a filter on the result while algorithm **B** performs a filter on the two RDDs and then joins the results. Assume the two algorithms obtain the same result. In general, which algorithm do you expect will cause less data shuffling?

☐ **A** ☐ **B**

7. Answer the following parts by typing in the spaces provided. Select from among the following words or phrases: "at most one," "multiple," "fast," "slow," "some," or "none."

(a) In a *narrow dependency*, each block of the parent RDD may be used by _____ block(s) of the child RDD.

Narrow dependencies are _____ since they require _____ of the data to be shuffled.

(b) In a *wide dependency*, each block of the parent RDD may be used by _____ block(s) of the child RDD.

Wide dependencies are _____ since they require _____ of the data to be shuffled.

8. (a) The *query optimizer* of Spark SQL is called

☐ Catalyst ☐ Cobalt ☐ Map Reduce ☐ Platinum ☐ Tungsten

(b) The *off-heap serializer* of Spark SQL is called

☐ Catalyst ☐ Cobalt ☐ Map Reduce ☐ Platinum ☐ Tungsten