CS 172 - Fall 2019 Quiz 1

Robert Arenas

TOTAL POINTS

8 / 13

QUESTION 1

1 2 / 3

- **0 pts** Correct
- **3 pts** Correct answer is (A), (B), (C).
- ✓ **- 1 pts** **Correct answer is (A), (B), (C). Missing 1 or selected (D)**
- **2 pts** Correct answer is (A), (B), (C). Missing 2 or selected (D)
- **2.5 pts** Correct answer is (A), (B), (C). Missing 2 and selected (D)

QUESTION 2

2 2 / 2

- ✓ **- 0 pts** **Correct**
- **3 pts** Correct answer is (a). The job is moved or run on the machines that store the data (if possible). Otherwise, we will alot of bandwidth transferring data over the network

QUESTION 3

3 0 / 2

- **0 pts** Correct
- ✓ **- 3 pts** **The correct answer is (a). The number of reducers are determined by cluster configuration or set by the developer for each MapReduce job.**

QUESTION 4

4 0 / 2

- **0 pts** Correct
- ✓ **- 3 pts** **False, That is the job of the 'Combiner' not the Partitioner.**
- **3 pts** No selection??

QUESTION 5

5 4 / 4

- ✓ **- 0 pts** **Correct**
- **2 pts** The Mapper receives input key/value pairs and applies some function on this chunk of data to generate key/value pairs as output. Then we have a sort and shuffle phase in between the Mapper and Reducer to route the key/value pairs to the right reducer. The reducers will receive a set of tuples for the same key to apply some aggregation function. (Answer must specifically addresses the sort and shuffle phase).
- **4 pts** No Answer
- **2 pts** The Mapper receives input key/value pairs and applies some function on this chunk of data to generate key/value pairs as output. Then we have a sort and shuffle phase in between the Mapper and Reducer to route the key/value pairs to the right reducer. The reducers will receive a set of tuples for the same key to apply some aggregation function. (Answer must specifically addresses the sort and shuffle phase).

UC Riverside

CS 172 : Information Retrieval

Quiz 1

Name: Robert Arenas Student ID: 862037366

1. The Hadoop Distributed File System (HDFS) stores the chunks (or blocks) of a large file across the DataNodes in the cluster. Which of the following is true? **(select all that apply)**

   (a) HDFS's replication policy is to store two copies on one rack, and a third copy on a remote rack.

   ● HDFS is a master-worker architecture composed of a NameNode and DataNodes.

   ● HDFS stores large files as blocks in a distributed manner.

   (d) HDFS is a peer-to-peer architecture where is there is no master node, only DataNodes.

2. The scheduler works with DFS and MapReduce(MR) to schedule jobs such that it exploits data locality. What did we mean by data locality? Consider how the jobs are assigned to the machines.

   ● It means the MR jobs are assigned to machines that already have the data

   (b) It means the data is moved to machines/nodes that will run the MR jobs

   (c) None of the above

3. How is the number of Reducers determined?

   (a) Use cluster configuration or value set by the user during Job creation

   ● Is determined based on the number of key-value pairs generated by the Mapper

   (c) Same way the number of Mappers are determined, based on the size of the input file.

   (d) Randomly generated number

4. In a MapReduce framework, a Partitioner is used to aggregate (key,value) pair at the Mapper so that it reduces the number of tuples sent to the Reducer. (TRUE) or FALSE

5. Describe the flow of a MapReduce job. What happens in-between the Map and Reduce?

   Data is recieved, Data is processed into key-value pairs by mappers, data is sorted into similar clusters then shuffled so that similar data goes to a reducer, data goes to a reducer, reducer produces output data.