Quiz 1

1. Stratified sampling is advantageous when the data distribution is not even.
   ( True / False )

2. All outliers within a dataset are invalid observations.
   ( True / False )

3. Consider a dataset comprising a set of (single value) observations. The average value for this dataset is 40 and the standard deviation is 10. If a reading has a value of 55, what would be the correct z-score for that reading?
   a. 1  b. 1.5  c. 2  d. 2.5

4. In MapReduce, a mapper must generate at least 1 key/value output pair for each input key/value pair.
   ( True / False )

5. In MapReduce, a reducer performs processing on a set of intermediate values that share a key.
   ( True / False )
Quiz 2
Design your map and reduce function to perform following data processing.

Find the 100 clients who spent the most electricity (kilowatt) for each zip code for the last month.

Files contain information about the last month only. The data is formatted as follows: `{customerID, TAB, address, TAB, zipcode, TAB, electricity usage, LINEFEED}`. Assume that each line will be used as the input to a Map function.

1. Which design pattern do you use to solve this problem?

   Top N

2. What is the output `<key, value>` pair for your map function?

   `<zipcode, [a list of top 100 info (info includes userID and electricity usage)]>`

3. What is the input `<key, a list of value>` pair of your reduce function?

   `<zipcode, [a list of all of the top 100 info (info includes userID and electricity usage)]>`

4. What is the major functionality of your reduce function?

   Select global top 100 within the zipcode and sort them based on the electricity usage

5. What is the output `<key, value>` pair for your reduce function?

   `<zipcode, [a list of global top 100 userIDs]>`
Quiz 3.
Suppose that you have two datasets USERS (2GB) and PIAZZA-POSTS (1.2TB). The dataset USERS maintains information about all users. USERS includes attributes such as userID, first name, last name, email, occupation, and date of birth. The dataset PIAZZA-POST stores all the posts that have appeared on the Piazza discussion board. PIAZZA-POST includes attributes such as, postID, userID, timestamp, post, and the type of the post.

Assume that you join (inner join) the datasets USERS with PIAZZA-POSTS on the foreign key of userID using MapReduce. To perform this join effectively, you use the replicated join pattern.

(1) To implement the replicated join pattern, we often use Hadoop’s DistributedCache. In this problem, which dataset should be uploaded to the DistributedCache?
   a. USERS only
   b. PIAZZA-POST only
   c. USERS and PIAZZA-POST
   d. a portion of the PIAZZA-POST

(2) The dataset uploaded to DistributedCache in problem (1) is available from the:
   a. namenode
   b. any remote node
   c. any mapper or reducer nodes within this MapReduce job
   d. any mapper or reducer nodes across the MapReduce jobs performed in the Hadoop cluster

(3) If a userID listed in USERS does not have any matching record in the PIAZZA-POST dataset, your output will include,
   a. Records from USERS with the corresponding userID and null values for the attributes in PIAZZA-POST
   b. Records from PIAZZA-POST with the corresponding userID and null values for the attributes in USERID
   c. Records from USERID with the corresponding userID and randomly selected records from PIAZZA-POST
   d. Your output will not include any record for this user ID

(4) To create a joined output, this MapReduce job does not require any reducer.
   ( True / False )

(5) For this MapProduce job, using a combiner may improve the performance.
   ( True / False )
Quiz 4.
Design your map and reduce functions to perform following data processing.

Consider a 5TB dataset that records financial transactions. You are required to use MapReduce to sort (in descending order) the entire set of records in this dataset based on the transaction amount (amountOfTransaction). **Design your MapReduce job to use “multiple” reducers to cope with the voluminous data.**

The data in individual records is formatted as follows: `{userID, TAB, date, TAB, transactionID, TAB, paymentMethod, TAB, amountOfTransaction, LINEFEED}`. Assume that each line will be used as the input to a Map function. You may use multiple MapReduce jobs to solve this problem.

Describe each MapReduce jobs with:
1. **Major functionality of your mapper**
2. **Input/Output of your mapper**
3. **Major functionality of your reducer**
4. **Input/Output of your reducer**

*Using TotalOrderPartitioner*

There are two phases: Analysis and Sorting phases
1. **Analysis phase (first MapReduce)**
   1. **Major functionality of mapper:** performing a simple random sampling
   2. **Input: each record**
      - **Output:** if the record is selected: `<amountOfTransaction, null>`
      - If the record is not selected: no output
   3. **Major functionality of reducer:** there will be a single reducer
      - No particular functionality. Return the keys only
   4. **Input `<amountOfTransaction, null>`**
      - **Output `<amountOfTransaction, null>`**

Separate task: Generate .seq file to specify ranges for partitioner using output from the “Analysis phase” job.

2. **Sorting phase (second MapReduce)**
   Uses TotalOrderPartitioner (or write your own customized partitioner) (Set this in driver)
   1. **This will be identity Mapper**
   2. **Input: each record, Output: `<amountOfTransaction, same as the input>`**
   3. **Major functionality of reducer:** writing the sorted values
      - No particular functionality.
(4) Input <the output of partitioner function for the amountOfTransaction, [a list of records]>
Output < the output of partitioner function for the amountOfTransaction, [a list of records]>
Quiz 5.

[Problem 1 ~ 4] Consider a MapReduce job running on a dataset of 1TB. This dataset contains multiple files stored in HDFS and the maximum size of each file is 1GB. This MapReduce job performs a random sampling algorithm over the data. Assume that the block size of HDFS is 128MB.

(1) If this MapReduce job does not specify the InputFormat, the maximum number of map tasks for this job is 1,000.

( True / false )

(2) If this MapReduce job does not specify the InputFormat, what is the minimum number of map tasks for this job?

a. 1,000  
   b. 1,116  
   c. 7,813  
   d. 8,000

(3) If you increase the size of InputSplit to 256MB, the number of map tasks for this job will be increased as well.

( True / False )

(4) If this MapReduce job uses TextInputFormat and the current InputSplit contains 15,000 linefeed signs in its text, what is the maximum number of inputs to the mapper in this task?

a. 15,000  
   b. 14,999  
   c. 15,001  
   d. 14,788

(5) A reducer of a MapReduce job retrieved 128 map outputs and merges these outputs. If the merge factor is 10, how many intermediate files will be created? Do not include the final merged file.

a. 13  
   b. 15  
   c. 10  
   d. 16
Question 1. In Hadoop’s YARN when a MapReduce job is created, the user’s jar file is passed to the Resource Manager as one of parameters.

(True/False)

Question 2. In Hadoop’s YARN implementation, the ResourceManager periodically pings the YarnChild to monitor the task.

(True/False)

Question 3. In Hadoop’s YARN framework, if the reducers have just completed the copy phase, what would be the progress estimated by YARN for the reduce phase?

a. $\frac{1}{3}$  
b. $\frac{2}{3}$  
c. $\frac{3}{6}$  
d. $\frac{5}{6}$

Question 4. The Hadoop’s combiner must be identical to the reducer of the job.

(True/False)

Question 5. In Hadoop’s YARN framework, a single partition is allowed to contain a group of different keys.

(True/False)