Quiz 1.  (Score: _______/1)
1. Stratified sampling is advantageous when sampling each stratum independently.
   ( 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 (Score: _________/1)

Design your map and reduce function to perform following data processing.
Find the client who spent the most electricity (kilowatt) for each zip code for the last month.
Assume that all the electricity usage measures are unique.
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 there are no duplicated values.

CASE 1) If you use a TOP10 pattern
(1) Specify the output <key, value> pair of your map function.
<zipcode, [a list of top 10 info (info includes userID and electricity usage)]>
(2) Specify the input <key, a list of [value]> pair of your reduce function.
<zipcode, [a list of lists those contain local top 10 info (info includes userID and electricity usage)]>
(3) What is the major functionality of your reduce function?
Go over the lists of local top 10 info and select global top 10 of the zipcode and sort them based on the electricity usage
(4) What is the output <key, value> pair for your reduce function?
<zipcode, [a list of global top 10 userIDs]>
(5) Design your combiner that improves the performance of this computation.
No combiner needed

CASE 2) If you use a Non-TOP10 pattern
(1) Specify the output <key, value> pair of your map function.
<zipcode, info (info includes userID and electricity usage)>
(2) Specify the input <key, a list of [value]> pair of your reduce function.
<zipcode, [a list of info (info includes userID and electricity usage)]>
(3) What is the major functionality of your reduce function?
Select global top 10 of the zipcode and sort them based on the electricity usage
(4) What is the output <key, value> pair for your reduce function?
<zipcode, [a list of global top 10 userIDs]>
(5) Design your combiner that improves the performance of this computation.
a) Specify the input of your combiner
<zipcode, [a list of info (info includes userID and electricity usage)]>
b) Specify the functionality of your combiner
Select the local top-10 of the given batch
c) Specify the output of your combiner
<zipcode, [a list of local top 10 info (info includes userID and electricity usage)]]>
Quiz 3. (Score: _______/1)
Suppose that you have two datasets USERS (1GB) and PIAZZA-POSTS (5.2TB). 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 userID

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

(5) For this MapReduce job, using a combiner may improve the performance.
   ( True / False )
Quiz 4. (Score: _______/1)
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(s) 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(s)
2. Input/Output of your mapper(s)
3. Major functionality of your reducer(s)
4. Input/Output of your reducer(s)

**ANSWER**

*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
   3. Output: if the record is selected: `<amountOfTransaction, null>`
      If the record is not selected: no output
   4. Major functionality of reducer: there will be a single reducer
      No particular functionality. Return the keys only

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 `<amountOfTransaction, [a list of records]>`
      Output `<amountOfTransaction, [a list of records]>`

Separate task: Generate .seq file to specify ranges for partitioning using output from the "Analysis phase" job.
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 Resource Manager periodically pings the YarnChild to monitor the task.  
( True / False )

Question 3. The Hadoop’s Combiner MUST have the same input and output format as that of the Mapper’s output.  
( True / False )

Question 4 and 5. Suppose that there is a MapReduce job running on an input file of size 2.56GB. The input file is stored in HDFS with a specified chunk (block) size of 128MB.

Question 4. If the user’s MapReduce job does not specify an input split, there will be 20 mapper tasks initialized to load the file.  
( True / False )

Question 5. If the start and end positions of a record is not fully encapsulated within a given block, the driver will load data blocks based on the start and end offset values of that record before the mapper task is assigned.  
( True / False )