**Quiz 6**

We design a MapReduce application that generates the top-10 most popular Twitter users (based on the number of followers) of a month based on the number of followers. Suppose that you have the following data file containing users' information regarding their followers. Assume that there are no duplicate records in this file. You should count the number of followers per user and find the top-10 most popular users in a result file with a single MapReduce run.

<table>
<thead>
<tr>
<th>UserID</th>
<th>FollowerID</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>A</td>
<td>K</td>
</tr>
<tr>
<td>B</td>
<td>X</td>
</tr>
<tr>
<td>B</td>
<td>Y</td>
</tr>
<tr>
<td>D</td>
<td>X</td>
</tr>
<tr>
<td>C</td>
<td>Y</td>
</tr>
<tr>
<td>A</td>
<td>X</td>
</tr>
</tbody>
</table>

1. Your mapper takes one record `<UserID, FollowerID>` as input. What would be the possible output pair of the map function to count the number of followers per user?
   a. `<FollowerID, {UserID, FollowerID}>`
   b. `<count, FollowerID>`
   c. `<FollowerID, userID>`
   d. `<userID, count>`

   where, `count` is a static variable that is defined as,
   ```java
   private final static IntWritable one = new IntWritable(1)
   ```

2. How many reducers do you need to complete this calculation?
   a. 64
   b. 1
   c. You should configure based on your available resource.
   d. You cannot complete this calculation with a single MapReduce run.

3. Consider a Combiner that tracks local counts of followers and emits only the local top10 users with their number of followers.

   This combiner will be able to improve the performance of the processing.
   (True / False)

4. This combiner will be able to generate the correct output.
   (True / False)

5. What could be a possible output of the combiner to generate correct results?
   a. `<FollowerID, {UserID, FollowerID}>`
   b. `<count, FollowerID>`
   c. `<FollowerID, userID>`
   d. `<userID, count>`

   where, `count` is a static variable that is defined as,
   ```java
   private final static IntWritable one = new IntWritable(1)
   ```
Quiz 7

[Question 1 - 4] Suppose that you should build a linear regression model to predict child’s weight when his/her height is given. You are using a dataset collected from 10,000 children.

\[ h_\theta(x) = \theta_0 + \theta_1 x \]

<table>
<thead>
<tr>
<th>ID</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>...</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height (Inches)</td>
<td>40</td>
<td>41</td>
<td>38</td>
<td>51</td>
<td>48</td>
<td>40</td>
<td>...</td>
<td>42</td>
</tr>
<tr>
<td>Weight (lbs)</td>
<td>38</td>
<td>39</td>
<td>36</td>
<td>42</td>
<td>41</td>
<td>40</td>
<td>...</td>
<td>45</td>
</tr>
</tbody>
</table>

Question 1. In above regression model, \( \theta_0 \) and \( \theta_1 \) are the parameter vectors.  
(True/False)

Question 2. With above regression model, the weight of child can be predicted as the value of \( h_\theta(x) \).  
(True/False)

[Question 3 and 4] To fit the linear regression model, assume that you perform the Batch Gradient Descent with MapReduce using 10 mappers.

For the given sample size 10,000, you will need to use following formula:

\[
\theta_0 \leftarrow \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})
\]

\[
\theta_1 \leftarrow \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})x^{(i)}
\]

Question 3. To calculate \( \theta_0 \), each mapper can take 1,000 data points, if they fit in a split with the default size.  
(True/False)

Question 4. To complete calculating \( \theta_0 \), this calculation will always require total 10 MapReduce jobs.  
(True/False)
Quiz 8
Suppose that there is a recommendation system for a music download service and they are using an Item-to-Item collaborative filtering algorithm to generate personalized recommendations. This algorithm uses a similarity matrix based on co-occurrences. Assume that there are 5 clients in this system and there are 10 songs in the service.
Clients: A, B, C, D, E
Songs: 01, 02, 03, 04, 05, 06, 07, 08, 09, 10

The system maintains the purchase history of all of the clients,
A: 01/02/03/09
B: 02/03/07/08
C: 09
D: 02/03/04/05/09/10
E: 08/09/10

This information can be represented using vectors such as,
P_A=(1,1,1,0,0,0,0,0,1,0)
P_B=(0,1,1,0,0,1,1,0,0)
P_C=(0,0,0,0,0,0,0,0,1,0)
P_D=(0,1,1,1,1,0,0,0,1,1)
P_E=(0,0,0,0,0,0,0,1,1,1)

Assume that there is no rating information available for these songs.

1) Complete the following co-occurrence based similarity matrix for the current purchase history. The value of the column M_{XY} is defined as the number of co-occurrences (co-purchases) of song X and Y. If X=Y, M_{XY} is the total number of occurrences of this song in the history of the purchase. (0.2 point)

<table>
<thead>
<tr>
<th></th>
<th>01</th>
<th>02</th>
<th>03</th>
<th>04</th>
<th>05</th>
<th>06</th>
<th>07</th>
<th>08</th>
<th>09</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>02</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>03</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>04</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>05</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>06</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>09</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>
(2) Assume that client E has logged-in and your system should generate recommendations for the client E based on the matrix generated in (1) and client E’s purchase history (If E has purchased any item, those items should not be included in your recommendation list.). Generate the top 3 recommendations using item-to-item collaborative filtering for the client E. Break ties arbitrarily, if there is a need to do so. (0.2 point)

2,3,7

(3) Design a general algorithm to provide recommendations using item-to-item collaborative filtering. Your algorithm must take client X’s purchase history vector and the similarity matrix generated in (1) as input parameters. Finally your algorithm must generate top k recommendations. Do NOT write Java code. Precise step-by-step instructions are required. Simple pseudo code is also acceptable. Break ties arbitrarily, if there is a need to do so. (0.2 point)

```plaintext
hashmap h
int k //top k
for item i purchased by user X
    for item t in the purchase history matrix M
        c = number of co-occurrence between i and t
        if c >0 and t is not a part of X’s purchased history
            put (t, c) into h

sort h on c
return first k
```
Quiz 9

(1) All of the write operations in BigTable write data in the main memory of the tablet server before it writes on the hard disk drive.  
(True/False)

(2) BigTable compresses data with two compression algorithms, BMDiff and Compressive Sensing algorithm.  
(True/False)

(3) Google’s BigTable performs a merging compaction when the memtable size reaches the specified threshold.  
(True/False)

(4) In Google’s Big Table, a tablet holds a contiguous range of rows.  
(True/False)

(5) Google’s BigTable uses a gossip protocol to manage the membership of storage nodes.  
(True/False)
**Quiz 10**

Suppose that there is a DHT ID circle with an identifier space of size $2^m$ where $m=3$. The DHT uses the Chord protocol and the ID-space spans: $0 \ldots (2^m-1)$. Initially, there is only one storage node A (id=4) on the identifier ring of a DHT.

Create a finger table for the node A. (Specify *.start*, *.interval*, and *.successor* for each entry)

<table>
<thead>
<tr>
<th>entry $i$</th>
<th>$\text{finger}(i).\text{start}$</th>
<th>$\text{finger}(i).\text{interval}$</th>
<th>$\text{finger}(i).\text{successor}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>[5 , 6 )</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>[6 , 0 )</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>[0 , 4 )</td>
<td>A</td>
</tr>
</tbody>
</table>
Quiz 11

1. For a skew join, Pig estimates the memory usage based on samples. (True/False)

2. Translating a US ZIP code using small unsorted lookup file can be implemented using merge join in Pig. (True/False)

3. To implement skew join, Pig must use the distributed cache to reduce the number of data loadings to the mappers. (True/False)

4. If Apache Pig’s JOIN operator includes a keyword “PARALLEL n”, Pig will perform the corresponding data joining process with n number of mappers and reducers. (True/False)