Queen Mary
University of London

Main Examination Period 2018

ECS640U **Big Data Processing** **Duration:** $2\frac{1}{2}$ **hours**

YOU ARE NOT PERMITTED TO READ THE CONTENTS OF THIS QUESTION PAPER UNTIL INSTRUCTED TO DO SO BY AN INVIGILATOR.

Instructions: This paper contains FOUR questions. **Answer ALL questions**.
Cross out any answers that you do not wish to be marked.

Calculators are not permitted in this examination.

Complete all rough workings in the answer book and cross through any work that is not to be assessed.

Possession of unauthorised material at any time when under examination conditions is an assessment offence and can lead to expulsion from QMUL. Check now to ensure you do not have any notes, mobile phones, smartwatches or unauthorised electronic devices on your person. If you do, raise your hand and give them to an invigilator immediately.

It is also an offence to have any writing of any kind on your person, including on your body. If you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms it will be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. A mobile phone that causes a disruption in the exam is also an assessment offence.

Exam papers must not be removed from the exam room.

Examiners: Felix Cuadrado and Arman Khouzani

Question 1

You have a dataset of tweets annotated with the personality of the author, according to the Myers Briggs Type Indicator (or MBTI for short). MBTI is a personality type system that divides everyone into 16 distinct personality types across 4 axis (introversion/intuition/-thinking/perceiving).

The dataset is a collection of rows, each one containing the following information for a single user:

```
userId;;;mbtiType;;;messages
```

The `messages` field contains the last 50 tweets the user posted (With each tweet separated by `;;;` (3 semicolon characters))

(a) Write a Map/Reduce program that computes the average length of the tweet messages from each personality type.

Use pseudocode for the program specification. You must clearly define the input and output of each one of your functions. State in your solutions any assumptions that are made as part of the program, as well as the behaviour of any custom function you deem necessary.

The code flow must be explained, discussing the input and ouput of each function that has been defined. You may use a diagram to illustrate the overall data flow.

[13 marks]

(b) Discuss how you would modify the program presented in 1a) in order to compute for each personality type both the number of members and the average length.

You should base your explanation in what information can be transferred in the flow of a Map/Reduce job.

[5 marks]

(c) Define a Combiner in the context of MapReduce jobs. Explain what would be the performance impact of adding a Combiner to the Map/Reduce jobs defined in 1a) or 1b). Discuss briefly the changes required to the code in order to add a Combiner.

[7 marks]

Question 2

(a) You have a dataset of user purchases for an online storefront such as Rakuten. The dataset contains one entry for each time a user clicked on a product from the online storefront, recording the following information [types in brackets]:

```
itemID[String],userID[String],category[String],action[String],time[Date]
```

`action` is a fixed String that can be either `"browse"`, when the user first clicks on the item, or `"buy"`, when the user wants to purchase the item.

This data is fed to the following MapReduce job. The keys are Strings with the unique id of the item accessed, and the values are ShopRequest objects with the full details (and methods to access each of the fields).

```
public void Map (String itemId, ShopRequest request) {
   if(request.getAction().equals("buy"){
      String category = request.getCategory();
      Date day = request.getTime().getDay();
      emit(new Pair(category,day), new Pair(itemId,1));
   }
}

public void Reduce (Pair key, List<Pair> values) {
   Hashtable<String,int> counts = new Hashtable();
   for(Pair value: values){
      String itemId = value.getLeft();
      int count = value.getRight();
      if(counts.containsKey(itemId)){
         int oldCount = counts.get(itemId);
         counts.put(itemId, oldCount + count);
      }
      else{
         counts.put(itemId, count);
      }

   }
   counts.sortInDescendingOrder().getFirst(10);
   for(Pair count: counts){
       emit(key, count);
   }
}
```

What will be the outcome of the Map/Reduce job? Explain the outcome as well as the nature of information exchanged in the job execution.

[10 marks]

(b) Suppose the program presented in 2a) will be executed on a dataset of 10 billion operations, collecting 30 days of data. 10% of the actions are “buy” actions, and 90% are “browse” actions. In total there are 100 different categories. The total input size is 2 Terabytes.

Which element of Hadoop is in charge of deciding how many resources will be used in order to execute this job?

Explain how many worker nodes will be involved in the job, assuming the cluster has a very large number of free nodes, and HDFS is configured with a block size of 128MB. State any assumptions you feel necessary when presenting your answer.

[9 marks]

(c) Explain the impact of data skew in parallel computing performance. Discuss whether the program presented in 2a) could experience data skew.

[6 marks]

Question 3

(a) Spark framework computations are based around RDDs.

(i) Define Spark RDD. Explain why Spark is an in-memory processing platform using the concept of RDDs.

(ii) Describe the lifecycle of an RDD by answering the following questions:

- How can a Spark program create new RDDs?
- When are RDDs created?
- How can RDDs be modified?
- How and when are RDDs destroyed by the Spark framework?

(iii) Give an example of an algorithm that takes advantage of Spark being an in-memory processing system.

[13 marks]

(b) Discuss the number of cluster nodes utilised by Spark in order to execute a transformation. Similarly, discuss how many nodes are needed to execute an action.

[6 marks]

(c) Discuss which transformation in Spark is more expensive to compute (in terms of both processing needs, and network usage) while handling similar size input: *flatMap* or *join*.

You can utilise MapReduce concepts to illustrate the difference between these operations. **[6 marks]**

Question 4

(a) This question is about data management in distributed processing systems.

(i) Explain the ’Move Computation to the data’ principle. Describe how HDFS and Hadoop support this principle.

(ii) Name the elements in charge of storing data in a Hadoop cluster and briefly list the role of each one.

(iii) Explain two mechanisms provided by HDFS to prevent data loss in a cluster.

[13 marks]

(b) This question is about Stream processing systems.

(i) Explain whether large-scale stream processing systems achieve low or high throughput, as well as low or high latency in comparison with batch processing systems.

(ii) Illustrate with an example the difference in output between a word count batch processing job, and a word count stream processing job, where both of them process tweet messages as its input data.

(iii) Explain the difference between pure stream processing, and the micro batch stream processing model. Present one advantage for each computation model.

[12 marks]

End of questions