Queen Mary
University of London

BSc Examination by course unit

xxth May 2016

ECS640 Big Data Processing Duration: 2 hours 30 minutes

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

Answer ALL FOUR questions

Calculators are not permitted in this examination. Please state on your answer book the name and type of machine used.

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 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: Félix Cuadrado, Graham White

Question 1 (25 marks)

a) The execution of a Map/Reduce job is completed through the coordinated execution of activities across a set of nodes in the cluster.

i. Explain how many cluster nodes will act as Mappers and Reducers for a given job. Make any assumption you consider necessary about the characteristics of the submitted job to the cluster.

[6 marks]

ii. During the execution of the previous Map/Reduce job, is there any type of information transfer between the worker nodes? If so, detail what types of nodes are the source and destination of each message flow.

[6 marks]

b) You have a dataset from TfL composed by Oyster card check-ins/outs into London tube stations. The format of the dataset is one line for each reading, following the format **`[String StationName,String Date,boolean isCheckIn,int cardNr].`**
An example reading would be*: `[Waterloo,201512131700,true,3345326554].`*
You have the following Map/Reduce program in pseudocode, which takes as input the just described TfL dataset.

```
Map (String stationId, Data checkInInfo)
Boolean isCheckIn =  checkInInfo.isCheckIn();
int day = checkInInfo.getDate().getDay()

if(isCheckIn){
     emit (new Pair(stationId, day), 1);
}
else{
     emit(new Pair(stationId,day), -1);
}
```

```
Reduce (Pair pair, List<int> checkIns)
int balance = 0;
for(int checkIn: checkIns){
    balance += checkIn;
}
emit (pair,balance)
```

Note that in the pseudocode a Pair object represents a tuple of two elements. The Boolean field isCheckIn has a true value when the user enters the station, and false when the user leaves the station.

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.

[8 marks]

c) Explain what would be the performance impact of adding a Combiner to the previous Map/Reduce job.

[5 marks]

Question 2 (25 marks)

a) You have a dataset with the September-December 2015 student activity on QMPlus. The data registers two types of user actions: file downloads, and upload of coursework submissions. Each interaction is recorded as a single entry. This data can be fed as input to a MapReduce job as a set of tuples (StudentId, Action). The keys are Strings with id of the student who interacted with QMplus, and the values are Actions with the full contents of the operation. Each Action has the following attributes:
`[ModuleID, Date, ActionType<download,submit>, TargetItemID]`

Design a Map/Reduce program (or a combination of Map/Reduce programs) that obtains for each module the week of the semester where most students submitted coursework assignments. In your pseudocode you can use a predefined method `getweek(date)` to parse the week of the semester.

[12 marks]

b) Suppose that there is a total of 1,000,000 student interactions over the collected period, recording the activity of 50,000 students at Queen Mary. QMplus has 500 different modules registered. Assume the job described in a) is executed in a cluster with 10 Mapper nodes, and 4 Reducer Nodes. Estimate how many key-value pairs will be emitted by each Mapper in this scenario. Estimate how many keys will be fed to each Reducer.

[7 marks]

c) Suppose that a program has a proportion *f* of its code which cannot be parallelised. Explain what Amdahl's law says about the maximum attainable speedup.

[6 marks]

Question 3 (25 marks)

a) Define the concepts of high availability and fault tolerance of distributed systems. How can distributed systems achieve fault tolerance? Describe one mechanism of Hadoop, and one feature of data centres that achieves fault tolerance at the platform and the infrastructure level respectively.

[9 marks]

b) How does load imbalance affect the performance of parallel computations? Name one possible cause of load imbalance occurring in Map/Reduce jobs.

[7 marks]

c) One of the advantages of pooling together a set of resources is to be able to share them more efficiently among multiple users/tasks. Describe the elements of Hadoop's architecture that cooperate to achieve efficient resource management and sharing. Illustrate their behaviour by explaining how Hadoop attends a new request to execute a job in the cluster.

[9 marks]

Question 4 (25 marks)

a) Define RDDs in the Spark framework. Explain how RDDs are created, and how RDD operations are parallelised by the Spark framework.

[9 marks]

b) Explain how both Map/Reduce and Spark support iterative computations, and compare the efficiency of both platforms when performing iterative computations.

[9 marks]

c) Explain the main difference between batch processing systems and stream processing systems from the input data point of view. Give one data analysis use case in which a stream processing system would be more appropriate (you have to specify both the nature of the data, and the analysis that should be computed on the data).

[7 marks]

End of Paper