Birkbeck

(University of London)

BSc/FD EXAMINATION

Department of Computer Science and Information Systems

Cloud Computing Concepts (BUCI028H6)

CREDIT VALUE: 15 credits

Date of examination: Thursday, 2nd June 2016
Duration of paper: 10:00am – 12:00pm (2 hours)

RUBRIC

1. *This paper contains* 5 *questions for a total of* 100 *marks.*
2. *Students should attempt to answer* **all** *of them.*
3. *This paper is not prior-disclosed.*
4. *The use of non-programmable electronic calculators is permitted.*

1. **(20 marks)**

Give brief answers to the following questions.

(a) What are the major service models of cloud computing? Which service models do Amazon EC2 and Amazon EMR belong to respectively? (5 marks)

(b) What is the main economic benefit of cloud computing? What do the three characteristics of big data — Volume, Velocity, and Variety — mean respectively? (5 marks)

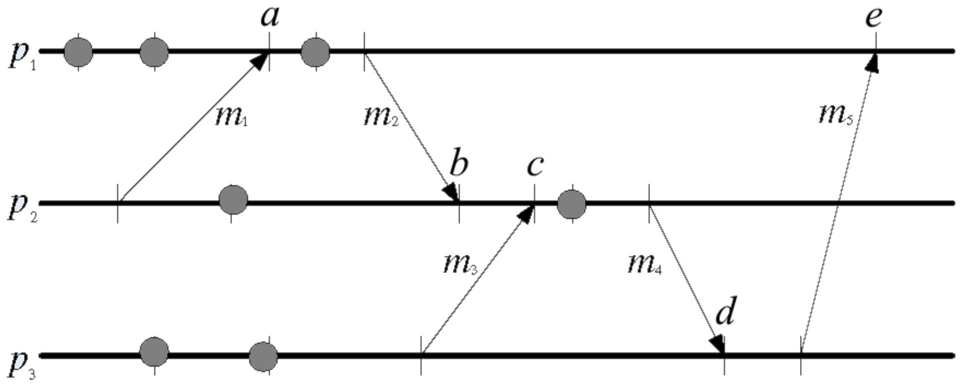
(c) What types of applications are most suitable to be put in the cloud? (5 marks)

(d) Why is it often more desirable to use a public cloud rather than a private cloud? (5 marks)

2. **(20 marks)**

Give brief answers to the following questions.

(a) What are the Lamport timestamps of the events a, b, c, d and e respectively in the following space-time diagram? (5 marks)



(b) What is a deadlock? What are the Coffman conditions for a deadlock to occur? (5 marks)

(c) Which three properties cannot be provided by a distributed system simultaneously, according to the CAP theorem? Which one of them do NoSQL databases usually compromise in favour of the other two? (5 marks)

(d) In RESTful APIs, what HTTP methods should be nullipotent and what HTTP methods should be idempotent? Which HTTP method should be used to replace a specific item in the collection? (5 marks)

3. **(20 marks)**

Give brief answers to the following questions.

(a) What are the pros and cons of the "in-mapper combining" design pattern compared with the combiner? (5 marks)

(b) What kind of reducer functions can be used directly as combiner functions? Provide three examples of such functions. Is it usually better to set the number of map tasks larger than the number of computer nodes in the cluster? Why? (5 marks)

(c) What are the three possible ways to perform a natural join of two relational tables using MapReduce? Which one of them has the fastest speed? Which one of them has the least restriction? (5 marks)

(d) What does RDD stand for in Spark? What does the lineage of an RDD mean? Why does an RDD need to remember its lineage all the time? (5 marks)

4. **(20 marks)**

There is a large text file of information about films stored in an HDFS over a number of machines. Each line of this file describes the details of one film in the following format.

$title \mid year \mid runtime \mid genres \mid stars$

The different fields are separated by the | character; the list of $genres$ and the list of $stars$ are both separated by commas; the $runtime$ is measured in minutes.

An example line is given below.

The Godfather | 1972 | 175 | Crime, Drama | Marlon Brando, Al Pacino, James Caan

You can assume that there are no duplicate records, and each distinct star's name is unique.

Write a MapReduce program (in pseudo-code) to calculate for each genre the average runtime of film in 1970s.

A combiner should be implemented to accelerate the computation.

The marking scheme is as follows.

(a) The projection (on the runtime) and selection (of films in 1970s) have been implemented correctly. (5 marks)

(b) The group-by (over genres) has been implemented correctly. (5 marks)

(c) The aggregation (to calculate the average) has been implemented correctly. (5 marks)

(d) The combiner has been implemented correctly. (5 marks)

Partial answers will be awarded marks accordingly.

5. **(20 marks)**

There is a large text file of information about films stored in an HDFS over a number of machines. Each line of this file describes the details of one film in the following format.

$title \mid year \mid runtime \mid genres \mid stars$

The different fields are separated by the | character; the list of $genres$ and the list of $stars$ are both separated by commas; the $runtime$ is measured in minutes.

An example line is given below.

The Godfather | 1972 | 175 | Crime, Drama | Marlon Brando, Al Pacino, James Caan

You can assume that there are no duplicate records, and each distinct star's name is unique. Write a MapReduce program (in pseudo-code), using the "stripes" pattern, to calculate for each star how many films he/she has co-starred with each other star (if they have co-starred before). For example, the output corresponding to the actor *Al Pacino* could be as follows.

Al Pacino — Marlon Brando: 1, James Caan: 1, Robert De Niro: 3

A combiner should be implemented to accelerate the computation.

The marking scheme is as follows.

(a) The mapper has been implemented correctly. (10 marks)

(b) The reducer and combiner have been implemented correctly. (10 marks)

Partial answers will be awarded marks accordingly.