Birkbeck

(University of London)

MSc EXAMINATION

Department of Computer Science and Information Systems

Cloud Computing (BUCI029H7)

CREDIT VALUE: 15 credits

Date of examination: Tuesday, 5th June 2018
Duration of paper: 2:30 pm – 4:30 pm (2 hours)

RUBRIC

1. *This paper contains* five *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.*

 ©Birkbeck College 2018

1. **(20 marks)**

Give brief answers (in a few sentences) to the following questions.

(a) A data mining program consists of four consecutive parts, P1, P2, P3 and P4 with the percentages of runtime being 10%, 30%, 40% and 20% respectively on a single processor. It is known that P1 and P3 can be parallelised while P2 and P4 cannot. If the problem/dataset size is fixed, how much speedup can this program achieve at most through parallel computing, according to Amdahl's Law? If the problem/dataset size can be arbitrarily large, how much speedup can this program achieve at most through parallel computing, according to Gustafson's Law? (5 marks)

(b) What does SPMD stand for? What does it mean in the context of parallel computing? Does it belong to data parallelism or task parallelism? (5 marks)

(c) What is a race condition? What is a deadlock? What is a livelock? (5 marks)

(d) What is eventual consistency? Why don't we insist on strong consistency in all distributed systems? (5 marks)

2. **(20 marks)**

Give brief answers (in a few sentences) to the following questions.

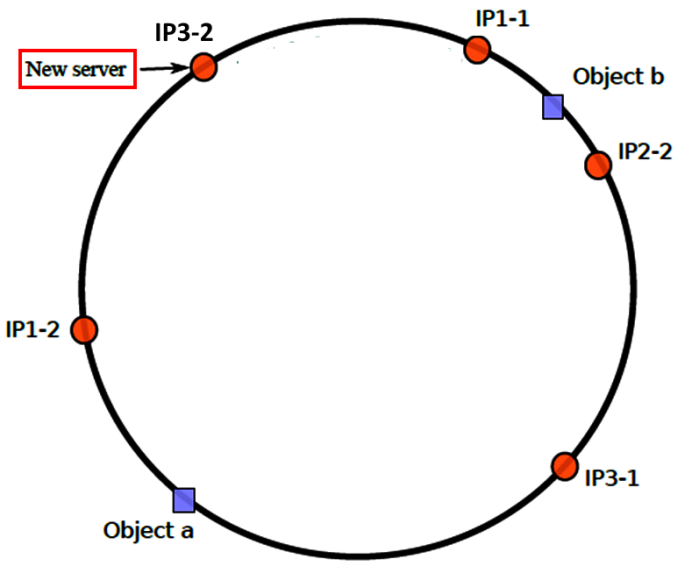
(a) What is the Chandy-Lamport snapshot algorithm used for? How many steps does it have? For a distributed system consisting of 4 processes that are fully connected with each other, how many messages would be exchanged to obtain a snapshot using this algorithm? (5 marks)

(b) In RESTful APIs, what HTTP methods should be nullipotent and what HTTP methods should be idempotent? Which HTTP method should be used to add new items to a collection at the given URI? (5 marks)

(c) What is the difference between the `map` operation in Spark and that in Hadoop? What is the difference between transformations and actions in Spark? What is the difference between narrow dependencies and wide dependencies in Spark? (5 marks)

(d) Why is it better to use consistent hashing rather than standard hashing for distributed indexing? Ignoring data replication (for failure recovery and load balancing etc.), in the following schematic diagram of consistent hashing, where should "object a" be stored, and what will happen when a new server "IP3-2" joins the cluster? (5 marks)

 ©Birkbeck College 2018



3. **(20 marks)**

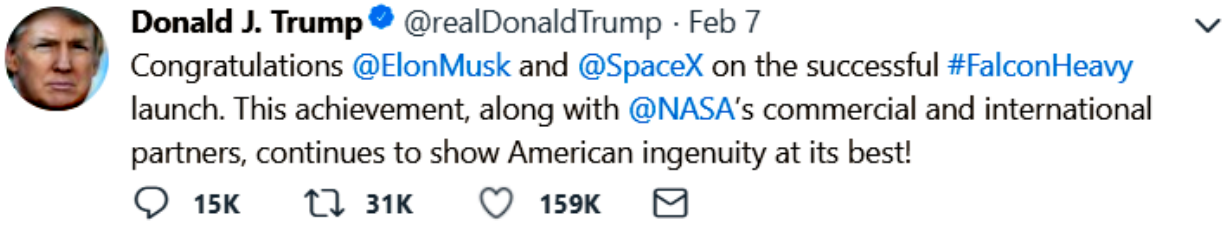
There is a large text file that contains all tweets from well-known politicians on Twitter since the beginning of 2018 up to now. It is stored in an HDFS over a number of machines. Each line of this file describes one tweet in the following format, where the different fields are separated by the | character (assuming that it does not occur in any tweet).

$$user \mid text \mid replies \mid retweets \mid likes \mid date$$

For example, the line

> realDonaldTrump | *Congratulations @ElonMusk and @SpaceX on the successful #FalconHeavy launch. This achievement, along with @NASAs commercial and international partners, continues to show American ingenuity at its best!* | 15000 | 31000 | 159000 | 07/02/2018

represents the tweet as shown in the following figure.



A tweet could contain hashtags (such as "*#FalconHeavy*") and mention usernames (such as "*@ElonMusk*") in its text.

Write a MapReduce program (in pseudo-code) to calculate for each hashtag the average number of retweets and the average number of likes in January 2018.
A combiner should be implemented to accelerate the computation.

 ©Birkbeck College 2018

4. **(20 marks)**

Consider the same large data file as described in the previous question.

Write a MapReduce program (in pseudo-code), using the "pairs" pattern, to calculate for each username the number of co-occurrences with another username if they have been mentioned in the same tweet before. For example, the output corresponding to the username *@ElonMusk* could be as follows.

@ElonMusk, @SpaceX: 300
@ElonMusk, @NASA: 200
@ElonMusk, @BillGates: 100

The "in-mapper combining" pattern should be implemented to accelerate the computation.

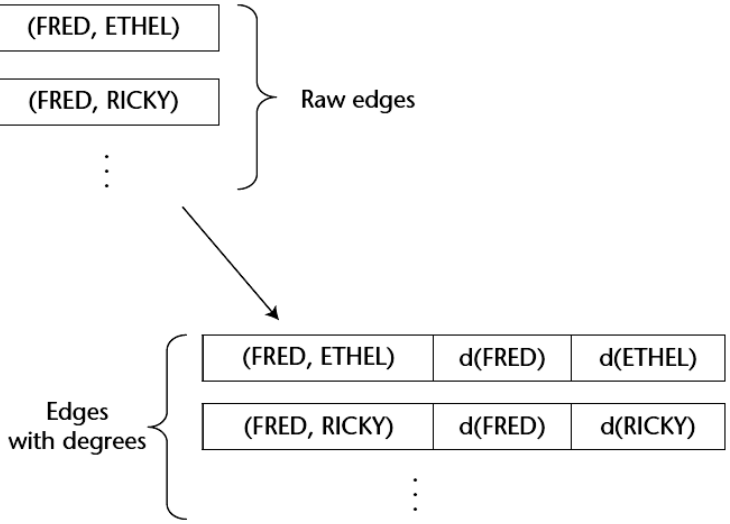
5. **(20 marks)**

Suppose that an *undirected* graph is stored as a file of edges (in the HDFS). Each line of the file is in the format "(u, v)" denoting that there is an edge between node u and node v. Each distinct edge in the graph occurs once and only once in the file. In other words, for an edge between node u and node v, the file contains either (u, v) or (v, u) but not both.

$[NB]$ The graph has too many nodes to be loaded into the memory of any single machine.
$[NB]$ It is not required to use combiners or in-mapper combining.
$[Tip]$ More than one MapReduce job could be used to accomplish the task.

(a) Write a MapReduce program (in pseudo-code) to find the node with the maximum degree (i.e., the number of edges). (10 marks)

(b) Write a MapReduce program (in pseudo-code) to augment each edge (u, v) with the node-degree information $d(u)$ and $d(v)$ which represent the degree of node u and the degree of node v respectively, as illustrated in the following figure. (10 marks)



 ©Birkbeck College 2018