COMP 4651 Cloud Computing & Big Data Systems

Fall Semester 2018

Final Examination Sample Questions

The final exam has the following six types of questions:

Question 1: Concept explanation

(a) What is cloud computing?

(b) What is virtualization?

(c) Why does GFS/HDFS employ a large block?

Question 2: Multiple choices

(1) The function of Secondary NameNode in HDFS is to

A. Serve as a backup for NameNode
B. Continue the functioning of NameNode
C. Serve as a checkpoint mechanism for primary NameNode
D. Provide advanced technology as compared with primary

(2) If we increase the size of files stored in HDFS without increasing the number of files, then the memory required by NameNode

A. Decreases
B. Increases
C. Remains unchanged
D. Cannot be decided

Question 3: What's the output of the following code?

data.txt:
Bob 23
Jimmy 32
Jason 27
Bob 26
Jason 75
Jimmy 24
Jason 45

```
data = sc.textFile("data.txt").map(lambda line: line.split(" "))
data.map(lambda x: (x[0], [int(x[1]), 1])) \
  .reduceByKey(lambda x, y: [x[0] + y[0], x[1] + y[1]]) \
  .map(lambda x: [x[0], x[1][0] / x[1][1]]) \
  .collect()
```

Question 4: Pseudo-code programming

Write a MapReduce pseudo-code to count the occurrence of each word in a text file:

Question 5: Debugging

The following code is inefficient. Explain why and how it can be fixed.

```
lines = sc.textFile("ReadMe.md", 4)
comments = lines.filter(isComment)
print lines.count(), comments.count()
```

Question 6: Code completion

Complete the following WordCount code:

```
text_file = sc.textFile("hdfs://...")
# Your code goes here:
counts =

counts.saveAsTextFile("hdfs://...")
```