ID2221 - Data Intensive Computing Platforms

24 October, 2018

1. **True/False questions.** Please briefly explain the reasoning behind your choice **(8 points)**

   (a) With MapReduce, each of the `R` reducers is responsible for producing $\frac{1}{R}$th of the amount of output data (true/false, why?)

   (b) GFS (HDFS) is used to store input, intermediate, and output files for MapReduce jobs (true/false, why?)

   (c) Assume we have a Dynamo storage with a hash function `H` that produces IDs between `0` and `32` to store and locate objects on the servers. If we have five servers with IDs `0, 3, 8, 18, 23` and `H(X) = 26`, then object `X` will be stored on the server with ID `23` (true/false, why?)

   (d) BigTable provides fault tolerance by replicating data on multiple tablet servers (true/false, why?)

---

2. **MapReduce question (8 points)**
   Suppose that you are given two documents with the following content:

   - `Document1: Hello world Hello Hadoop`
   - `Document2: Hello Spark`

   We want to generate a list of locations (i.e., word number in the document and identifier for the document) for each word occurrence. The output generated by your program should look like:
   `Hello` → `Document1 : 1, 3 | Document2 : 1`
   `world` → `Document1 : 2`
   `Hadoop` → `Document1 : 4`
   `Spark` → `Document2 : 2`
   Write out in pseudo-code the steps taken in Hadoop's map and reduce phases to generate the above output. Please also specify the input and output of the `map()` and `reduce()` functions. Assume the identifier for each document is provided as the key to the `map()` function.

3. **Spark questions (8 points)**

   (a) Draw the lineage graph for the following code and mention which connections are narrow and which ones are wide?

```
val a = sc.parallelize(...)
val c = sc.parallelize(...)
val e = sc.parallelize(...)
val b = a.groupby(...)
val d = c.map(...)
val f = d.union(e)
val g = b.join(f)
```

   (b) Explain briefly how does spark join two tables, if

      i. both tables are so big that none of them can be loaded into memory of one computer

      ii. one table is big and the other one is small, such that only the small one can be loaded in memory of one computer.

---

4. **Streaming questions (8 points)**

   (a) Assume you want to use Storm to implement the word count application. Explain what spouts and bolts you need, and what types of grouping do you use between them.

   (b) Briefly compare the fault tolerance model of Storm, Spark Streaming and Flink.

---

5. **Graph questions (8 points)**

   (a) What is the difference between message passing and shared memory models in graph processing platforms?

   (b) Assume we have a graph, in which all vertices have a local numeric value. Write a vertex-centric Gather-Apply-Scatter pseudo-code to update the local value of the vertices with the minimum value in the graph. For example, if a graph has three vertices `A, B`, and `C`, with values `4, 2` and `5`, respectively, we would like to end up with value `2` at all vertices.