Name of the student: Matriculation Number:

Large-scale Data Processing on the Cloud

Fall 2014

Date: 04-11-2014

Total duration: 2 Hrs Total Marks: 100

1. Explain short and long term implications of cloud. (6 Marks)
2. Please explain major themes (at least 5) of cloud computing. (6 Marks)
3. Please explain the basic models of scaling information systems. (4 Marks)
4. Explain Map and Fold functions in Functional programming and provide the pseudo code for calculating the sum of the cubes of the elements in a list. (6 Marks)
5. Please explain the partition and combine functions of MapReduce in detail. (6 Marks)
6. Please explain in detail the MapReduce algorithm for generating Inverted Index. (4 Marks)
7. Please explain in detail how a MapReduce job runs in a Hadoop framework, describing its components (8 Marks)
8. Please explain the pairs and stripes examples discussed in the lecture and explain how synchronization is achieved in both the cases in solving the conditional probabilities problem. Explain conditional probability equation and give MapReduce algorithms of both approaches. (10 Marks)
9. Elaborate TF-IDF Term Weighting algorithm, discussed in information retrieval lecture and explain how you can compute it using MapReduce. (10 Marks)
10. Describe each of the Pig data structures (Relation, Bag, Tuple, Field) used in pig. What are their differences from typical relational database table structure components are what can each of them contain? (8 Marks)
11. Which of the two languages - Pig or Hive - would be more suitable when you need to rewrite an existing application containing more than 200 SQL for deployment in Hadoop cluster? Why is your choice better? (6 Marks)
12. Describe 3 main differences between Spark and Hadoop MapReduce. What needs to be considered when choosing which one to use for a large scale data processing task? (6 Marks)
13. Please explain the differences between relational and NoSQL databases (4 Marks)
14. Explain the four non-relational data models with examples. (8 Marks)
15. What are the differences between sampling and sketching? Describe the sequential reservoir sampling algorithm. Name two applications where sampling or sketching techniques are used. (8 Marks)