Big Data Analytics

Exercise Sheet 5

Prof. Dr.Dr. Lars Schmidt-Thieme, Mohsan Jameel

Information Systems and Machine Learning Lab University of Hildesheim

May 26th, 2016

Submission until June 1st, 2016, to mohsan.jameel@ismll.de

Exercise 1: Distributed File System (8 points)

a) Define Hadoop Distributed File System (HDFS) architecture and explain the role of each component? What are the difference between HDFS and Google File System (GFS)?

b) What is a MapReduce Framework? What are the key features of each component and define their role. Also list the benefits of programming with MapReduce?

Additional reference to the lecture slides:

http://www.cse.buffalo.edu/faculty/bina/MapReduce/HDFS.ppt

http://rakaposhi.eas.asu.edu/cse494/notes/s07-map-reduce.ppt

http://www.dcs.bbk.ac.uk/~dell/teaching/cc/paper/cacm08-01/p107-dean.pdf

Exercise 2: Apache Hadoop (HDFS and MapReduce (7 points)

a) Install Apache Hadoop on your local machine (standalone version) and run the Word Count example. You have to add screen shot of the output of word count program.

b) For the files given below, count occurrence of each fruit using MapReduce framework. You have to show each step in the process i.e. what are the key-value pair, what happen at Map phase, Reduce Phase and final output (no need to write code or algorithm). You can assume three mapper and two reducers for this example

Input files:

a. Content of file on Node 1 (apple, orange, mango, apple, banana)
b. Content of file on Node 2 (apple, apple, plum, kiwi, kiwi, mango, mango)
c. Content of file on Node 3 (orange, orange, plum, grapes, kiwi, mango, apple)

Reference link:

https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

Exercise 3: Getting familiar with cluster (2 points bonus points)

Follow the instruction given in the tutorial 5, log in to the cluster and run HelloWorld program. When you submit job provide screenshot of output of qsub, qstat and output of your program. Provide qsub command (or script) that you used to submit job.