TECHNISCHE UNIVERSITÄT DARMSTADT

TK1: Distributed Systems - Programming & Algorithms

4. Programming Assignment
Submission Date: 14.01.2014

Prof. Dr. Max Mühlhäuser
Dr. Immanuel Schweizer

Jens Heuschkel, MSc.
Michael Stein, MSc.

TELEKOOPERATION
Fachbereich Informatik
Hochschulstr. 10
64289 Darmstadt

If you have questions regarding this Assignment, please contact our expert "Kunal Saxena" (kunal1991@hotmail.com).

Task 1

a) Setup a distributed Hadoop Cluster of at least 3 Physical Nodes.
b) Run the Multi node cluster and load the dataset into HDFS.

Task 2

a) Compile eclipse plugin for Hadoop.
b) Configure environment for MapReduce Programming in Eclipse.

Configuration files for both above tasks are provided

Task 3

We have a Data set (NASDAQ Exchange Daily 1970-2010 Open, Close, High, Low and Volume) which consists of stock related information for the given time frame. The data is in CSV (comma separated values) file containing fields which are self-explanatory. A skeleton for MapReduce programming will be provided which consist of all the API's which are required.

Compute the Following:

a) Perform "word count" on the given Dataset. (Dataset should be present in the HDFS, not on the local machine)
b) Write a MapReduce Program to count the total volume done for each stock.

EXPECTED RESULTS:

1. Provide a screenshot when the HDFS is started on the distributed Hadoop cluster.
(Check with $jps for the Daemons running in the cluster)
NOTE: Simulate the cluster with at least 1 name node and 3 data node.
(FOR the groups of less than 4 people, HINT: name node can act as data node too)

2. Upload the dataset in the HDFS within a directory named usr/tucanid/username.
Traverse the HDFS in the WebUI and provide a screenshot for the above directory.

3. Provide the screenshot of the output for the word count problem. (Results should be present in the HDFS, traverse through WebUI)

4. Provide the screenshot of the output for the Stock_volume problem (results should be present in the HDFS, traverse through WebUI)

5. Provide the source code for both tasks (code should be properly commented).