Assignment 5

For your convenience, we will work with a virtual machine containing the relevant environment for Hadoop, and the systems for the next assignment sheets. You can download the VMWare image from here: the Moodle course.

The virtual machine has Ubuntu 20.04 as operating system. The user credentials are:

- Username: `de`
- Password: `de`

Once you login, you will find Java 8, HDFS, Yarn, Eclipse already installed and configured. You can find Eclipse in the Downloads directory; the workspace to use is scala-eclipse-workspace.

Exercise 1 – Your first Hadoop task

Once you boot the machine and login, you will find a README file for the Hadoop assignment sheet in the desktop. Open it and follow the instructions to start HDFS and Yarn. Be patient: you may have to wait a few seconds for the startup operations to complete.

Once HDFS and Yarn are up and running, you can check that everything is working by opening Firefox. The browser is already configured to open tabs with the relevant Web interfaces.

Following the instructions in the README file, copy file.txt from the home directory of user de into the /input directory in HDFS. Check that the copy operation is successful with the HDFS commands.

Now you can start Eclipse and open the provided FirstMapReduce project. Inspect the code and explain what it does.

The MapReduce project has already been compiled into a .jar file, which you can find in the home directory of user de. Nevertheless, you should try to export the project yourself into a new .jar file and include it in your submission. Then, follow the instructions in the README file and execute it. After execution, check the output file on HDFS using “ls” and “cat”. Include the created files in your Moodle upload.

Exercise 2 – Your own Hadoop job

Create a copy of the wordcount project with the name "gradesStatistics". Change the code such that the Hadoop job computes the

a) average grade of each student,
b) average grade of each course,
c) best grade given in each course
d) average grade of all students of each examinator.

from the input file "examResults.txt" on Moodle.

Execute each job and include the result files as 2a, 2b, 2c, 2d in your Moodle upload.

Hints: The file "examResults.txt" follows the Hadoop "KeyValueTextInputFormat" format, where the key is the course id and the value is a json string. For json parsing you might use `org.json.simple.parser.JSONParser.`

Further Readings:

- https://hadoop.apache.org/docs/r2.9.0/api/org/apache/hadoop/mapred/KeyValueTextInputFormat.html
- https://code.google.com/archive/p/json-simple/