Middle East Technical University
Department of Computer Engineering

CEng 495
Cloud Computing
Spring 2019-2020 Homework 3

Due Date: 23.05.2020, 23:55

This homework aims to get you familiar with MapReduce paradigm. You are going to develop and deploy a MapReduce application by using Apache Hadoop Packages and Java language.

Keywords: Cloud Computing, Hadoop, Apache, MapReduce, Java

1. Apache Hadoop

- Download and install the latest stable release of Apache Hadoop.
- You must have the required JDK version to use Hadoop.
- See the useful links section.

2. Specifications

- You will implement a Java code with Hadoop environment to analyze the input files consisting of student grades on various courses.
- You will be given a folder containing input text files.
- The inputs will have the following form:

<STUDENT_ID > <COURSE_ID> <GRADE>

- GRADE can be any integer value between 1 and 100.
- A student can only have one grade from a course. (i.e. no duplicate STUDENT_ID COURSE_ID pairs will be given as input)
- If a student has a GRADE greater than or equal to 60 on a course, it means the student has a passing grade from that course. Otherwise, he/she fails that course.

- Your program will execute the following tasks:

  a. List the number of students taking for each course. (**cap**)
  b. List the number of courses passed for each student. (**pass**)
  c. List the average grade for each student. (**avg**)
  d. For each course, list the average grade of students who have passing grade, and the average grade of failing students. i.e. passing average on “part-r-00000”, and failing average on “part-r-00001”. (**twolist**)

- For simplicity, you can assume STUDENT_ID and COURSE_ID have the same standard like METU ID’s: 7 digit numbers
- The outputs of MapReduce are sorted according to the keys by default, thus you do not need to change anything for the order of the outputs.
- For question b, do not give output with a student ID if he/she has no passing grade. (i.e. there should not be something like “1440142 0” on the output file. This rule also holds for question d, since the average of ‘none’ is not defined.)
- There can be more than one input file. Your program should read all the files in the input folder.
- You can see the input and output formats on the sample input and output files. Since black-box testing will be used for grading, be sure to stick to the format.
- Your code must be in Java language using the Apache Hadoop library.
- Your codes will be evaluated automatically in Local (Standalone) Mode of Hadoop. Assuming that all of the Java files of your solution exist in the current directory, the command sequence below will be executed in order to build the solution:

```
hadoop com.sun.tools.javac.Main *.java

jar cf Hw3.jar *.class
```

- The output jar file will be tested with commands given below with different inputs.

```
hadoop jar Hw3.jar Hw3 cap input output_c

hadoop jar Hw3.jar Hw3 pass input output_p

hadoop jar Hw3.jar Hw3 avg input output_a

hadoop jar Hw3.jar Hw3 twolist input output_t
```

3. Useful Links

- Apache Hadoop: http://hadoop.apache.org/
- To download: http://kozyatagi.mirror.guzel.net.tr/apache/hadoop/common/stable/
- Install guide: https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-common/SingleCluster.html#Installing_Software (Note that the most common problem is to forget to set the environment variables on file "hadoop-env.sh")
- You can look at the following tutorial and use the corresponding code as a base for your work: https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

4. Submission

- In this assignment, you are expected to submit your Java code(s) to ODTÜClass. For submission on ODTÜClass, a tar.gz archive file (named hw3.tar.gz) that contains all your source code files.
- The work you submit should be implemented by only you and genuine.
- We have zero tolerance policy for cheating. There is no teaming up! People involved in cheating will be punished according to the university regulations and will get 0. You can discuss design choices or language preferences, but sharing code between each other or submitting third party code as a whole is strictly forbidden. In case a match is found, this will be considered as cheating.