Middle East Technical University
Department of Computer Engineering

CENG 495
Cloud Computing
Spring 2020-2021 Homework 3

Due Date: 20.06.2020, 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 travel duration information between cities.
- You will be given a folder containing input text files.
- The inputs will have the following form:

<CITY_1 > <CITY_2> <DURATION>

- DURATION can be any integer value more than or equal to 1.
- For simplicity, you can assume CITY_1 and CITY_2 will be strings consisting of lowercase ASCII characters.
- The order of cities (being city 1 or city 2) is not important. No inputs with city 1= city 2 will be given.

- Different inputs can contain exactly same entries duplicated. You will treat them as different entries.
- Your program will execute the following tasks:

  a. List the total travel duration. (**tot**)
  b. List the number of occurrences for each city in the data. (**city**)
  c. List the average duration for each route. (**avg**)
  d. Separate the travel data according to the duration of the routes and list the number of occurrences for each route. First, you will partition the data into 4 parts. i.e. the routes that takes less than or equal to 5 hours, or $dur \leq 5$ will be listed on file "part-r-00000", the routes that have duration: $5 < dur \leq 10$ will be listed on file "part-r-00001", $10 < dur \leq 15$ will be on "part-r-00002" and $15 < dur$ will be on "part-r-00003". (**sep**)

- Precedence is not important for the input but it is important for the output. i.e. input "ankara istanbul 6" will be treated same as input "istanbul ankara 6" but the ROUTE will be outputted as "ankara-istanbul".
- ROUTE will be a string containing CITY-CITY where cities will be ordered in terms of their characters (alphabetic precedence).
- 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 d, do not give output with a ROUTE if it does not have any entry within the required interval for a partition (i.e. there should not be something like "ardahan-edirne 0" on the "part-r-00000" output file unless there is a travel data between these cities that takes less than or equal time of 5 hours.
- Some routes may have different durations within different entries. So, they will be counted on different partitions for question d. (See "ankara-izmir" on sample input files.)
- 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 tot input output_t

hadoop jar Hw3.jar Hw3 city input output_c

hadoop jar Hw3.jar Hw3 avg input output_a

hadoop jar Hw3.jar Hw3 sep input output_s
```

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.