Middle East Technical University | Department of Computer Engineering

CENG 495

Cloud Computing

Spring '2016-2017

Assignment 3 - MapReduce

Due date: May 28th 2017, Sunday, 23:55

1 Objective

This homework aims to get you familiar with Map Reduce paradigm using Hadoop environment.

In short; Develop a MapReduce application by using Apache Hadoop Packages, Java language, and Eclipse IDE).

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

2 Problem Definition

3 Specifications

1. In this homework you will write a Java code with Hadoop environment for analizing and finding some information about a sequence of random numbers.
2. You will be given an input text file which contains a random sequence of numbers.
3. Your program will conduct the following tasks:
   (a) Finding the number of **occurences** of each number in the input file.
   (b) Finding the **max valued number** of "Even" and "Odd" numbers in the input file.
   (c) Finding the **average** of "Even" and "Odd" numbers in the input file.
4. Your program will print the results of these tasks.

3.1 Document Format

3.1.1 Input Format

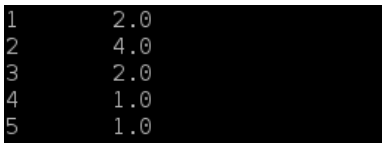
1. Every line respresents the random sequence of numbers.
2. No negative or floating numbers is going to be tested.
3. Seperator of the two numbers is only going to be blank aka "␣" character.

An input sample

3.1.2 Output Format

1. Output formats of each are different. Below you can find three different formats:

First task sample output



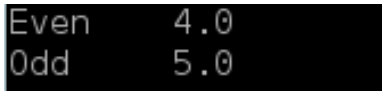
[Number] [Occurences]

[Number] [Occurences]

[Number] [Occurences]

[Number] [Occurences]

Second task sample output



Even [Max value of even]

Odd [Max value of odd]

Third task sample output

Even [Average value of even]

Odd [Average value of odd]

3.2 Sample Input and Output

3.2.1 Example

1. Suppose the following random sequence of numbers is given as an input:

   1 2 3 4 5

   1 3

   2 2 2

2. First task should find all of the occurences of each number:

   1 2.0

   2 4.0

   3 2.0

   4 1.0

   5 1.0

3. Second task should find the maximum value of both 'Even' and 'Odd' numbers (Since 5 is the maximum of odds and 4 is the maximum of evens, program should print them):

   Even 4.0

   Odd 5.0

4. Last task should find the average value of both 'Even' and 'Odd' numbers:

   Even 2.4

   Odd 2.6

3.3 Program Specifications

1. The solution must be in Java language using the Apache Hadoop library.
2. Your solutions 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

3. The output jar file will be tested with commands given below with various different input files values.

   hadoop jar hw3.jar Assignment3 count input.txt outputtask1

   hadoop jar hw3.jar Assignment3 max input.txt outputtask2

   hadoop jar hw3.jar Assignment3 average input.txt outputtask3

4 Deliverables

1. You are expected to submit an archive file hw3.tar.gz including all of your source code files.

5 Evaluation

1. Submission schedule will be strict. However you will have ***3*** extra days with penalty. Penalty formula is ***-5xDayxDay***. Submission time will be determined by your last source code submission to coursepage on COW.
2. Your application will be tested only once except the objection. You should also keep in mind that you should attend the objection hour of the day which you submit your code if you want to object your grade. Any objection after that day will not be accepted.
3. You can visit A301 for objections.

6 Useful Links

1. You can download and set up Hadoop on a single node from here.
2. Another useful tutorial for installation and configuration can be found from here.
3. You can access Hadoop tutorial from here.
4. An important problem about hadoop solution can be found from here.
5. Apache Hadoop 2.7.3v download link can be found from here

7 Cheating

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 using third party code is strictly forbidden. In case a match is found, this will be considered as cheating. Even if you take a "part" of the code from somewhere/somebody else - this is also cheating. Please be aware that there are "very advanced tools" that detect if two codes are similar. So please don't think you can get away with by changing a code obtained from another source. Also, we will check your codes with last years' students' solutions, so beware about that!