Distributed Systems Project, Spring 2014 – Assignment 2

IMPORTANT!

If you have not yet done so, provide your department computer user ID by email to Liang.Wang@cs.helsinki.fi so that we can create a directory for you.

Assignment

In this assignment, you are supposed to use Hadoop to analyze a large data set. Hadoop is an open source implementation of the MapReduce paradigm which we have seen briefly in the course. We provide the environment where to run Hadoop and also the data set. The data set is simply a set of numbers in a text file. See below for the exact format.

Requirements

You need write a program that uses Hadoop to provide an answer to the following questions:

1. What are the minimum and maximum values, the average value and the variance?
2. What is the value of the median of the data set?
3. Answer the questions under point 1, but restrict the data set to values above the value of the first quartile.

For the first question, you will need to provide 4 numbers, 1 for the second, and 4 for the third.

Documentation

In the documentation, you should explain how your code solves the problem and how it uses Hadoop. You also need to provide the answers to the above questions.

Grading

Grading is based on the correctness of the program and the answers, quality of the program code, and associated documentation.

Guidelines

The assignment is individual work. You can of course discuss any problems you encounter with other students, but sharing code is not allowed and if found, will be considered as plagiarism.

You are free to choose any programming language.

Deliverables

Program source code with documentation. The document should explain how you have solved the problems and provide answers to the 3 questions from Requirements section.

Timeline

The assignment is due on January 28th at 10:00. No extensions will be given.

Return

Return your code and documentation by email to Liang.Wang@cs.helsinki.fi as one tar-archive. Please indicate clearly your name and student ID in every file.

Set Up

Hadoop (version 2.2.0) has been installed on the Ukko cluster, in CoNe group folder. The complete path is /group/home/cone/hadoop. Here is a very brief instruction to help you start quickly. Additional help will be provided in the Q&A sessions as needed.

0. Provide your user ID by email to Liang.Wang@cs.helsinki.fi so that we can create a directory for you in the Hadoop installation.

1. First, you need log into melkki with the following command:

```
ssh username@melkki.cs.helsinki.fi
```

2. In order to use our Hadoop installation, you need to add the search path to your shell environment by adding the following line into .bashrc file. You can find .bashrc file in your home directory.

```
# Hadoop
export GPH=/group/home/cone
export PATH=$PATH:$GPH/hadoop:$GPH/hadoop/bin:$GPH/hadoop/sbin
```

3. Then logout and re-login into melkki and your new shell environment should work now. Try the following commands to see if Hadoop HDFS works:

```
hadoop fs -du .
hadoop fs -help
```

4. Download the example from the course webpage, and run compile.sh. If you can see the following output, it means Hadoop is correctly running for you now.

```
minValue        10.25634718
```

You may also see a lot of other diagnostic output, but these 5 lines should be almost the last things you see.

5. The data set file is in the following format.

```
0     11839923.64831265
1     5710431.90800272
2     2638393.35244932
3     1317095.15852382
4     23746643.73449028
5     3807727.13565876
6     5441016.03486564
```

```
7     7578570.44049178
8     5188792.46384957
9     20315369.89568381
```

The first field is an index and the second field is the value that you are supposed to use in the assignment. The file has about 536 million lines.

More Information

You may also find the following links useful:

Hadoop HDFS shell command:
https://hadoop.apache.org/docs/current2/hadoop-project-dist/hadoop-common/FileSystemShell.html

Hadoop commands:
https://hadoop.apache.org/docs/current2/hadoop-project-dist/hadoop-common/CommandsManual.html

For those who want to install Hadoop on their own machines, please refer to the following link. Note there is a significant change in Hadoop architecture in 2.2.0. Make sure the tutorials you find on the Internet match the version you are using.
https://hadoop.apache.org/docs/current2/index.html

Note that the data set is only available on our Hadoop installation, so making your own installation is only helpful in getting to know Hadoop better. You will **not** be able to do the assignment on your own installation.