Class 1: Homework

Realtime and Big Data Analytics

Fall 2018

Homework

Class 1

1. Please read in Hadoop: The Definitive Guide, fourth edition:

   - Chapter 1: All.
   - Chapter 2: Java and non-Java programmers, read up to page 30.

     If you program in a language other than Java, you must use Hadoop Streaming. Read middle of p.37 through p. 41. Please post on the forum if any problems.

2. Please read “MapReduce: Simplified Data Processing on Large Clusters”, Dean and Ghemawat, OSDI 2004.

   http://static.usenix.org/event/osdi04/tech/full_papers/dean/dean.pdf

   - Sections 5, 6, 7 are optional.
   - Please write a brief summary of this paper – one paragraph.

Homework

Class 1

3. You will require access to a Hadoop system in order to complete the homework assignments for this class. Choose one of these options:

a. ***Easiest***: NYU's Hadoop Cluster, Dumbo

There is an NYU HPC Hadoop cluster (Dumbo) available for homework and projects. Please contact me or the TAs for more information on obtaining an account - this is a newly updated cluster (hardware and software). Even if you will use the Quickstart VM, I recommend getting an account on Dumbo. The NYU HPC IT team provides support. I recommend option **a.** above, but if your host machine has too little memory, Dumbo is your next best choice. There is no cost for using Dumbo. If you experience problems, the TAs can help and so can the NYU HPC IT group at: hpc@nyu.edu

b. Download the Cloudera Quickstart VM (or Hortonworks Sandbox)

Everything you need in order to complete the Hadoop homework assignments is already installed in the VM. Download from here - Virtual Box format is highly recommended because Virtual Box has very good terms on their free trial license:

https://www.cloudera.com/downloads/quickstart_vms/5-10.html

* *Please make sure your host machine meets the memory requirements posted on the VM website.*

c. You may also opt to create your own cloud-based Hadoop clusters. This costs money, so not a good alternative for homework but possible for projects if there are free trials available (AWS and GCP have free trials).

d. You may opt to install a Hadoop distribution (cost-free) available from several vendors on your Mac or on a Linux box/VM. Let me or the TA(s) know if you are taking this path because this is the hardest of all and you may encounter difficulties.

This will work in a Linux-based OS, but in the past, doing this on Windows was futile. That said, you may be able to find a Hortonworks distribution that you can try and it may work (they recently put some effort into a Windows distro in collaboration with Microsoft).

Homework

Class 1

4. Once you have your Hadoop environment established, you can run the simple MapReduce example in the book – it's the weather dataset example which is part of your reading assignment.

Detailed homework instructions:

a. Try out the Hadoop HDFS commands in your Hadoop environment, you will need this to do the assignment.

Try issuing these commands:

```
hdfs dfs -ls /              -- To see the contents of the top-level directory in HDFS
hdfs dfs –ls                -- To see the contents of your user directory
hdfs dfs –mkdir myNewDir    -- To create a new directory named ‘myNewDir’ in your user directory
hdfs dfs –ls                -- To verify that you now have a directory called ‘myNewDir’
hdfs dfs –rm –r myNewDir    -- To remove directory ‘myNewDir’
hdfs dfs –ls                -- To verify that you have successfully removed the directory called ‘myNewDir’
```

A great reference is here.

b. Read pp.17-27.

The MapReduce program that I would like you to run is in the book in Example 2-3, 2-4, and 2-5 (pp.22-26) - you don't have to write your own program, just use the book example.

The weather data that you must use is in the book example (just 5 lines in a file) - see middle of page 23.
You will need to **pad out the '…'** in the sample data with dummy data, **or change the indexes** in the program to make this work.

c. Type in the program and input the data as shown in Example 2-3, 2-4, and 2-5 in the book, run your program:

```
hadoop jar yourJarFile.jar className </path/to/your/input/data/directory> </path/to/your/output/data/directory>
```

(If this doesn't work, let us know.)
Be sure to use the data shown in the book - you may have to adjust the indexes in the code or pad out the data to match the indexes.

d. Upload homework to NYU Classes. To receive full credit, please hand in all of the following items:

- Your source code files, small sample input, and job output (similar to the output in section ‘A Test Run’ on pp.25-26)
- Evidence that the program ran successfully (**e.g. screen shots and/or output log**)
- Evidence that the correct output is obtained

e. Please use the Forum on NYU Classes if you experience any difficulties. The TAs and I will help you get your environment working.