


Module 2 Homework



New York University
Courant Institute of Mathematical Sciences

Homework

1. If you have not already done so, please obtain an HPC account (you will need it to complete homework assignments and to complete the project).

For help, use the Forum or send an email to me and our TA(s).

2. Please complete last week's TDG reading if you haven't already done so.

3. This week, please read:

- Chapter 2: Page 30-37 (Scaling Out, until Streaming)
- Chapter 3: Stop at top of p.48 (HDFS Federation), read p. 70-71 (Network Topology and Hadoop), read bottom of p.73-top of p.76 (Replica Placement until Parallel Copying with distcp).

4. Please read: "The Google File System", by Ghemawat, Gobioff, and Leung.

Link: <http://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf>

Read sections 1 and 2 at least. You will notice a difference in terminology when compared with HDFS.

Please summarize the paper in one paragraph.

Homework

5. Run the MaxTemperature example from the book

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:

<code>hdfs dfs -ls /</code>	-- To see the contents of the top-level directory in HDFS
<code>hdfs dfs -ls</code>	-- To see the contents of your user directory
<code>hdfs dfs -mkdir myNewDir</code>	-- To create a new directory named 'myNewDir' in your user directory
<code>hdfs dfs -ls</code>	-- To verify that you now have a directory called 'myNewDir'
<code>hdfs dfs -rm -r myNewDir</code>	-- To remove directory 'myNewDir'
<code>hdfs dfs -ls</code>	-- 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. **screenshot(s)**)
- 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.