Lab 1: Introduction and Setup
Big Data Spring 2016

The goal of this lab is to make sure you have accounts on both NYU's HPC cluster and Amazon's AWS, and can log into these systems and run sample jobs. These will be the two primary computing systems we will use in this course.

Before starting this lab, you will need:

1. Your HPC account username and password. You should have received an email with these.
2. Your AWS account username and password. You should have already applied for these by following the instructions under "Setting up your AWS account" at http://www.vistrails.org/index.php/AWS_Setup.
3. SSH/SCP tools
    a. If you are using MacOS, you will run your commands from the terminal (under Utilities) which already has the programs you need installed. You do not need to install any additional software.
    b. If you are using Windows, you will need to download PuTTY, PSCP, and PuTTYgen from http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html. (Click putty.exe, pscp.exe, puttygen.exe, to download the files).
4. A basic familiarity with using the linux command line. There are many resources available online. For example, see http://freeengineer.org/learnUNIXin10minutes.html for a tutorial, and http://ryanstutorials.net/linuxtutorial/cheatsheet.php for a cheatsheet of commands. You might want to keep a resource like this open in another tab while you work.

Part 1: Accessing HPC and Setup

Logging in to the cluster:

1. Log into the main HPC node. To do this,
    a. On MacOS, open the terminal and type `ssh your_netid@hpc.nyu.edu`
    b. On Windows, open PuTTY. In the "Host Name" field, type `your_netid@hpc.nyu.edu`, and then click "Open" at the bottom.
2. Enter your password when prompted.
3. From the HPC node, log into the Hadoop cluster. To do this, type `ssh dumbo`. Enter password again (if prompted).

You will be using a set of commands, and it will save you some time to first create aliases for them.

Create command aliases:

1. Once you've logged in to "dumbo", make sure you are using the "bash" shell. The prompt should say something like "-bash-4.1$". If not, type "echo $0" and it should say "-bash". If it says something else, like "tsch", type "bash" to switch to bash. (You are welcome to use another shell, but you'll need to modify the following instructions, so we don't recommend it unless you really know what you're doing.)
2. Once on "dumbo", run the following commands on your terminal (Note: you should not have any spaces around "=" signs!):

```
alias hfs='/usr/bin/hadoop fs '
export HAS=/opt/cloudera/parcels/CDH-5.4.5-1.cdh5.4.5.p0.7/jars
export HSJ=hadoop-streaming-2.6.0-cdh5.4.5.jar
alias hjs='/usr/bin/hadoop jar $HAS/$HSJ'
```

3. To be able to re-use these aliases every time you login to dumbo, append the following lines to the end of your .bashrc file (full path: /home/your-netID/.bashrc):

```
alias hfs='/usr/bin/hadoop fs '
export HAS=/opt/cloudera/parcels/CDH-5.4.5-1.cdh5.4.5.p0.7/jars
export HSJ=hadoop-streaming-2.6.0-cdh5.4.5.jar
alias hjs='/usr/bin/hadoop jar $HAS/$HSJ'
```

   To edit a file from the command line, you can use any of the built-in text editors --- vi, emacs, or nano --- for example, type `vi .bashrc` (in your home directory). Of these three editors, nano is probably the easiest to learn, but has less advanced functionality. Emacs and vi have more advanced features and will take a few weeks to master. There are many tutorials and how-tos for these programs online (google for them).

4. Now type

   ```
   source .bashrc
   ```

   to create the aliases. Bash 'sources' .bashrc automatically at login, so you won't have to type this command again.

Part 2: Running Hadoop on HPC

1. Move the files map.py, reduce.py, wikipedia.txt from the lab resources to the HPC cluster.
   a. On MaxOS, from the terminal, type `scp file_location/file_name your_netid@dumbo.es.its.nyu.edu:/home/your_netid/.`
   b. On Windows, run cmd.exe. Navigate to the folder where you saved pscp.exe. Type `pscp file_location/file_name your_netid@dumbo.es.its.nyu.edu:/home/your_netid/.`
2. Log in to the Hadoop cluster dumbo, following instructions above.

3. From dumbo, you will now copy the data file to HDFS. Type `hfs -copyFromLocal /home/your_netid/wikipedia.txt wikipedia.txt`
4. Check if the file is on HDFS. Type `hfs -ls` and you should see the file wikipedia.txt.
5. To run the job, type `hjs -file map.py -mapper map.py -file reduce.py -reducer reduce.py -input /user/your_netid/wikipedia.txt -output /user/your_netid/wikipedia.output`
6. The outputs of this job are now in HDFS, in the directory user/your_netid/wikipedia.output.
7. Copy the files to your local directory by typing `hfs -get /user/your_netid/wikipedia.output  output`
8. Navigate to the folder output (by typing `cd output`) and you can view the output of the program (by typing `cat name_of_output_file`)

Part 3: Running Spark on HPC

Spark allows you to write and run applications quickly in Java, Scala, Python and R.

1. Move the files wordcount.py and sample.txt from the lab resources to the HPC cluster, as in step 1 of the previous section.
2. Log in to the Hadoop cluster dumbo, following instructions above as before.
3. Type `hfs -copyFromLocal sample.txt sample.txt`
4. Type `spark-submit --num-executors <10-100> wordcount.py sample.txt >> sparkoutput.txt`
    a. Above, replace the "<10-100>" with a number between 10 and 100, e.g., `spark-submit --num-executors 100 wordcount.py sample.txt >> sparkoutput.txt`
    b. Note that you may see warnings while running the job; ignore these for now.
5. To view the output of the job, type `cat sparkoutput.txt`

Part 4: Set up a Key Pair

Generate a public/private key pair:

1. Follow the instructions at: http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-key-pairs.html#having-ec2-create-your-key-pair
    a. For windows users, you need to convert your .pem file to a .ppk file. Follow instructions under "Converting Your Private Key Using PuTTYgen" on http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/putty.html). (We won't need this today, but may for future labs.)

Part 5: Running a Hadoop Job on AWS

First, we create an S3 bucket for input/output storage. (S3 is Amazon's cloud storage.)

1. Go to https://console.aws.amazon.com/s3 .
2. Click "Create Bucket".
3. Call the bucket something unique, e.g., "your_netID-ds1004-sp16". (It doesn't really matter what you call it -- it just has to be unique across Amazon S3.)
4. This takes you to the "All Buckets" page of your S3 Management Console; click the name of the bucket you just created.
5. Using the Create Folder button, create a folder in your bucket called "wordcount".
6. Inside "wordcount", create subfolders "input" and "logs".
7. Upload map.py and reduce.py to the wordcount folder
8. Upload sample.txt to the wordcount/input folder

Now we fire up an Amazon EMR (Elastic MapReduce) cluster

9. Go to https://console.aws.amazon.com/elasticmapreduce
10. Click Create Cluster
11. Use defaults, except for
    - Cluster name: wordcount
    - Logging S3 folder: this should be
      s3://bucket/wordcount/logs
      where bucket is the name of the bucket you just created
    - Number of instances: 1
    - EC2 key pair: select name of key pair you created in earlier step
11. After the cluster has started running, click Add Step.
12. Use defaults, except for:
    - Step type: Streaming program
    - Name: wordcount
    - Mapper: s3://bucket/wordcount/map.py
    - Reducer: s3://bucket/wordcount/reducer.py
    - Input S3 location: s3://bucket/wordcount/input
    - Output S3 location: s3://bucket/wordcount/output
13. Click the "Add" button.
14. Wait for the EMR job to finish (this can take up to 10 minutes).
15. Navigate to your S3 bucket and look in the newly-created output folder; the output of the wordcount program is stored in the files here.

Part 6: Terminating your EMR cluster

You MUST terminate your cluster when you are finished with it; simply closing the browser does not terminate your instance. If you do not terminate, it will continue to run and use up your AWS credits, and will then start to **charge your credit card** once your credits run out!.

Terminating an instance:

1. Go to https://console.aws.amazon.com/elasticmapreduce/
2. Select your cluster in the list.
3. Click the “Terminate” button.
4. Click the “Terminate” button in the pop-up window.
5. Click the “Yes, Terminate” button.
6. After a few minutes, navigate to https://console.aws.amazon.com/ec2/v2/ and verify that you show “0 Running Instances”.