**Exercise 8**

*More Apache Spark and Python, EC2*

**Prior Knowledge**

Unix Command Line Shell

Simple Python

**Learning Objectives**

Using Spark on EC2

Accessing S3 files on Spark

Reading CSV files in Spark

Seeing the differences between Spark and Hadoop by performing the Wind Analysis in Spark

Spark SQL

**Software Requirements**

(see separate document for installation of these)

* EC2 credentials
* Flintrock

**Part A. Starting Spark in EC2**

1. There is a project from the creators of Spark to run it in EC2, but it is not very good! Instead we will use a tool called **flintrock**
2. Before we can use flintrock, you need to modify the config file for flintrock so that it uses your own keys. Edit the flintrock config file:

subl ~/.config/flintrock/config.yaml

It will look something like:

The source for this is here:

<https://freo.me/flintrock-conf>

This is modified from the original in a couple of ways. Firstly, it gives the Ireland region and AMI files. Secondly, there is an “instance-profile-name”. This is a AWS feature that gives the running VM access to other APIs - in this case S3.

1. Change the key name and identity file to match your key name and identity file.
2. Make sure   
     
   install-hdfs: False
3. Make num-slaves: **2**
4. Save the file
5. You should now be able to launch a cluster in Amazon:  
     
   flintrock launch oxcloXX-sc   
     
   (using your XX)
6. Now you should see something like:   
   Ignore the Apache mirror warning.   
     
   If you have issues you can try:

flintrock --debug launch oxcloXX-sc

1. Let’s login to the master (all one line):  
     
   flintrock login oxcloXX-sc  
     
   You see something like:  
   
2. This basically just SSH’s you into the master. You could do the same from the EC2 console as before.
3. Now we will start pyspark once again but this time from the flintrock SSH session.  
     
   This time we are going to add in a Spark Package that supports accessing S3 data (Amazon object storage). **Once again, all one line**  
   pyspark --master spark://0.0.0.0:7077  
    --packages org.apache.hadoop:hadoop-aws:2.7.4
4. You should see a lot of logging, eventually ending with:  
     
   
5. It is perfectly possible to get Jupyter to talk to Spark on our cluster, but it is slightly complex, so we will just use the normal Python command-line for the moment.
6. We are going to use Spark’s SQL support, which in turn uses Apache Hive.
7. This combined with the CSV package we saw earlier makes it very easy to work with data.   
   First let’s tell spark we are using SQL. In the Python command-line type:  
     
   from pyspark.sql import SQLContext  
   sqlc = SQLContext(sc)
8. Now let’s load the data into a DataFrame. (one line)  
     
   df = sqlc.read.csv('s3a://oxclo-wind/2015/\*',header='true', inferSchema='true')

Spark should go away and think a bit, and also show some ephemeral log lines about the staging.   
  
Ignore the warning:  
  
ObjectStore:568 - Failed to get database default, returning NoSuchObjectException

1. The df object we have is not an RDD, but instead a DataFrame. This is basically a SQL construct. (But we can easily convert it into an RDD as you will find out shortly). It is similar to the Pandas dataframe (and convertible into one: <https://docs.databricks.com/spark/latest/spark-sql/spark-pandas.html>)
2. We can print a nice table showing the first few rows with:  
     
   df.show(4)  
   (I shrunk this so you can see the table nicely!)  
   
3. We can also convert the DataFrame into an RDD, allowing us to do functional programming on it (map/reduce/etc)  
     
   winds = df.rdd
4. Let’s do the normal step of mapping the data into a simple <K,V> pair. Each column in the row can be accessed by the syntax e.g. row.Station\_ID  
     
   We can therefore map our RDD with the following:   
   mapped = winds.map(lambda s: (s.Station\_ID, s.Wind\_Velocity\_Mtr\_Sec))
5. We can simply calculate the maximum values with this reducer:  
     
   maxes = mapped.reduceByKey(lambda a, b: a if (a>b) else b)
6. And once again collect / print:  
     
   for (k,v) in maxes.collect(): print k,v  
     
   Because python uses indentation, it can’t tell if this is the end of the statement so you will see:  
     
   ...

Press Enter.

1. You will see a bunch of log before the following appears:  
   
2. You can also turn the response of a collect into a Python Map, which is handy. Try this:  
     
   maxes.collectAsMap()['SF04']
3. You can also try:  
   print maxes.collectAsMap()

**PART B – Getting Jupyter running with Flintrock**

1. Quit the pyspark REPL (Ctrl-D) and get back to the ec2 command line
2. Type the following commands to install and run jupyter into your master node (available here: <https://freo.me/flintrock-j>)   
   
3. You will see something like:  
   
4. Don’t try to access that URL just yet. That is a URL that is only accessible from within the master node running on EC2 at the moment.
5. To allow us to access that URL, we need to setup an SSH tunnel to the master node.

Start a new Ubuntu terminal window.   
Find the name of the master node once again:

flintrock describe oxcloXX-sc

Now start ssh thus (all one line, and replace the hostname)

ssh -i ~/keys/oxcloXX.pem -4 -fN -L 8888:localhost:8888

ec2-user@ec2-34-244-248-67.eu-west-1.compute.amazonaws.com

Explanation  
*-4 - use IPv4 only  
-fN - go into the background and don’t execute any  
 remote command  
-L xxxx:localhost:yyyy   
 port forward from xxxx on the local   
 server to yyyy on the remote server*

1. Now we can open that URL in the other window. You are now accessing the Jupyter server running in EC2. Now you can use the Jupyter model as before.
2. *Note that any python code you save here will be stored on the AWS instance and deleted when you destroy the cluster!*  
     
   **PART C - SQL**
3. There is an easier way to do all this if you are willing to write some SQL.
4. We need to recreate the DataFrame first, so run this in a cell:  
     
   from pyspark.sql import SQLContext

sqlc = SQLContext(sc)

df = sqlc.read.csv('s3a://oxclo-wind/2015/\*',header='true',

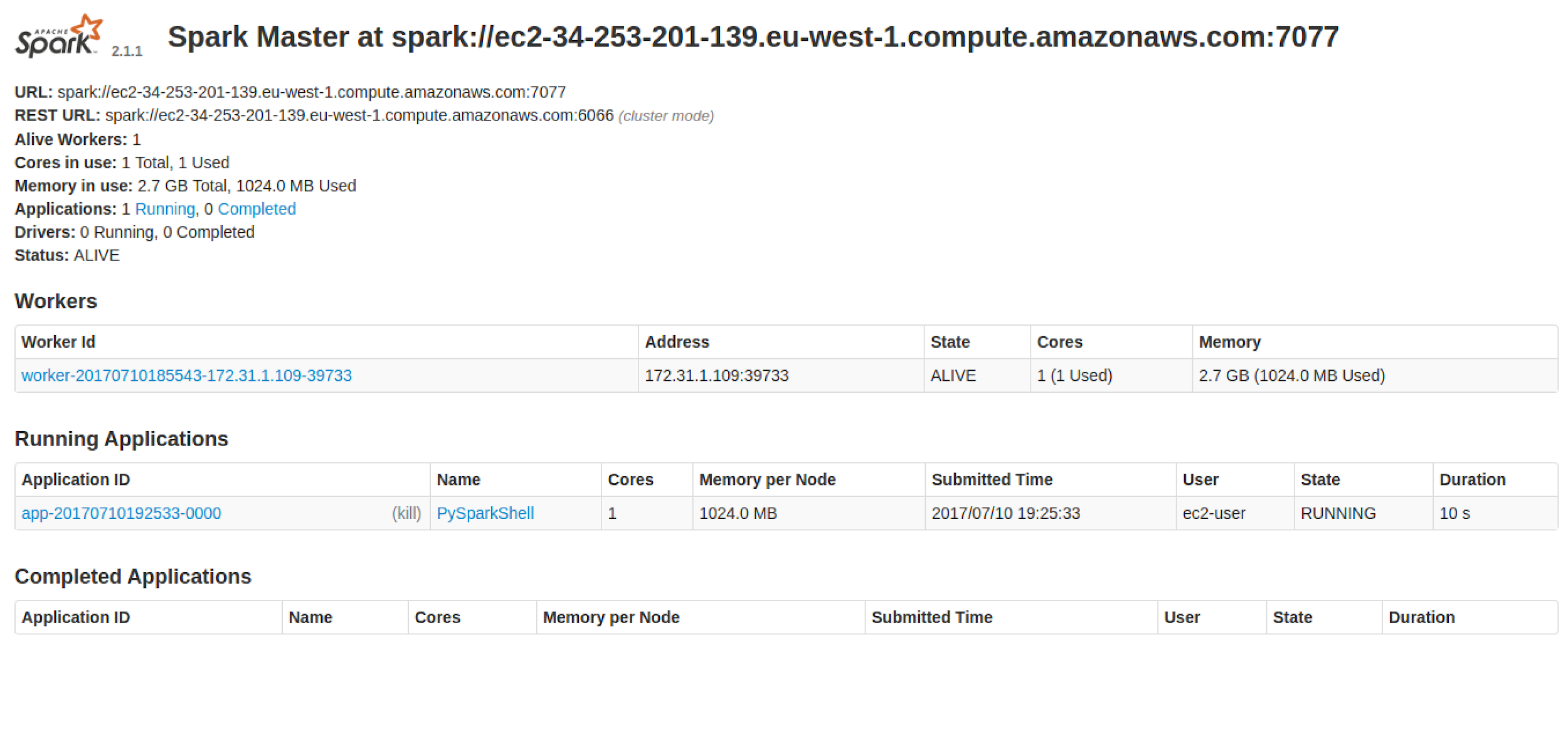
inferSchema='true')

df.show(4)

1. Now we need to give our DataFrame a table name:  
     
   df.registerTempTable('wind')
2. Now we can use a simple SQL statement against our data.   
     
   sqlc.sql("SELECT Station\_ID, avg(Wind\_Velocity\_Mtr\_Sec) as avg,max(Wind\_Velocity\_Mtr\_Sec) as max from wind group by Station\_ID").show()
3. Bingo you should see:  
   
4. Recap. So far we have:
   1. Started Spark in EC2
   2. Loaded data from S3
   3. Used SQL to read in CSV files
   4. Explored Map/Reduce on those CSV files
   5. Used SQL to query the data.
5. Find the IP address of the Spark Master: in your Ubuntu start a new terminal and type:  
     
   flintrock describe oxclo01-sc

You should see something like:  
  


1. Go to e.g.   
   http://ec2-52-214-61-215.eu-west-1.compute.amazonaws.com:8080   
     
   using the master’s DNS address (not the one in this text)  
   You should see something like:



1. If you want you can try adding another slave and then rerun the analysis. You can see the extra core working in the Web UI

flintrock add-slaves --num-slaves 1 oxcloXX-sc

If you need it the code is here:  
<https://freo.me/wind-sql>

1. We must remember to stop our cluster as well (its costing money…)  
   From Ubuntu terminal   
     
   flintrock destroy oxcloXX-sc  
     
   Type y when prompted.
2. Congratulations, this lab is complete.