**Exercise 7**

*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 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 (except with more lines):

Launching 3 instances...

[54.154.17.100] SSH online.

[54.154.17.100] Configuring ephemeral storage...

[54.154.17.100] Installing Java 1.8...

[34.253.201.139] SSH online.

[34.253.201.139] Configuring ephemeral storage...

[34.253.201.139] Installing Java 1.8...

[54.154.17.100] Installing Spark...

[34.253.201.139] Installing Spark...

[34.253.201.139] Configuring Spark master...

Spark Health Report:

\* Master: ALIVE

\* Workers: 1

\* Cores: 1

\* Memory: 2.7 GB

launch finished in 0:03:49.

1. Let’s login to the master (all one line):  
     
   flintrock login oxcloXX-sc  
     
   You see something like:

Warning: Permanently added '34.253.201.139' (ECDSA) to the list of known hosts.

Last login: Mon Jul 10 18:55:35 2017 from host109-156-251-208.range109-156.btcentralplus.com

\_\_| \_\_|\_ )

\_| ( / Amazon Linux AMI

\_\_\_|\\_\_\_|\_\_\_|

https://aws.amazon.com/amazon-linux-ami/2017.03-release-notes/

1 package(s) needed for security, out of 1 available

Run "sudo yum update" to apply all updates.

[ec2-user@ip-172-31-6-32 ~]$

1. This basically just SSH’s you into the master. You could do the same from the EC2 console as before.
2. Now start pyspark once again.   
     
   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 com.amazonaws:aws-java-sdk-pom:1.10.34,org.apache.hadoop:hadoop-aws:2.7.2
3. You should see a lot of logging, eventually ending with:

---------------------------------------------------------------------

| | modules || artifacts |

| conf | number| search|dwnlded|evicted|| number|dwnlded|

---------------------------------------------------------------------

| default | 71 | 71 | 71 | 0 || 70 | 70 |

---------------------------------------------------------------------

:: retrieving :: org.apache.spark#spark-submit-parent

confs: [default]

70 artifacts copied, 0 already retrieved (36388kB/279ms)

Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties

Setting default log level to "WARN".

To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).

17/07/10 18:58:16 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

17/07/10 18:58:38 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is not enabled so recording the schema version 1.2.0

17/07/10 18:58:38 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException

17/07/10 18:58:41 WARN ObjectStore: Failed to get database global\_temp, returning NoSuchObjectException

Welcome to

\_\_\_\_ \_\_

/ \_\_/\_\_ \_\_\_ \_\_\_\_\_/ /\_\_

\_\ \/ \_ \/ \_ `/ \_\_/ '\_/

/\_\_ / .\_\_/\\_,\_/\_/ /\_/\\_\ version 2.1.1

/\_/

Using Python version 2.7.12 (default, Sep 1 2016 22:14:00)

SparkSession available as 'spark'.

>>>

1. 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.
2. We are going to use Spark’s SQL support which in turn uses Apache Hive.
3. 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  
     
   sqlContext = SQLContext(sc)
4. Now let’s load the data into a DataFrame. (one line)  
     
   df = sqlContext.read.format('com.databricks.spark.csv').  
   options(header='true', inferschema='true').  
   load('s3a://oxclo-wind/2015/\*')
5. Spark should go away and think a bit, and also show some ephemeral log lines about the staging.
6. 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.
7. 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!)

+----------+--------------------+--------------+----------------+-----------------+---------------------+---------------------------+------------------+-------------------------+----------------------------+

|Station\_ID| Station\_Name|Location\_Label|Interval\_Minutes|Interval\_End\_Time|Wind\_Velocity\_Mtr\_Sec|Wind\_Direction\_Variance\_Deg|Wind\_Direction\_Deg|Ambient\_Temperature\_Deg\_C|Global\_Horizontal\_Irradiance|

+----------+--------------------+--------------+----------------+-----------------+---------------------+---------------------------+------------------+-------------------------+----------------------------+

| SF15|Warnerville Switc...| Warnerville| 5| 2015-01-5? 00:05| 1.628| 8.1| 148.5| 0.92| 0.061|

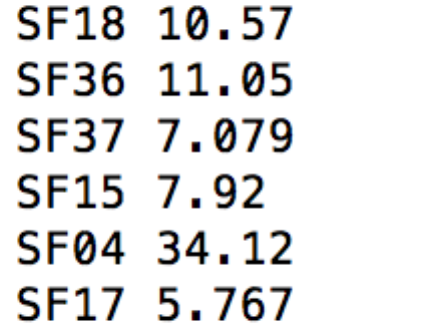
| SF15|Warnerville Switc...| Warnerville| 5| 2015-01-5? 00:10| 1.519| 9.4| 151.1| 0.717| 0.064|

| SF15|Warnerville Switc...| Warnerville| 5| 2015-01-5? 00:15| 1.482| 8.7| 142.7| 0.627| 0.059|

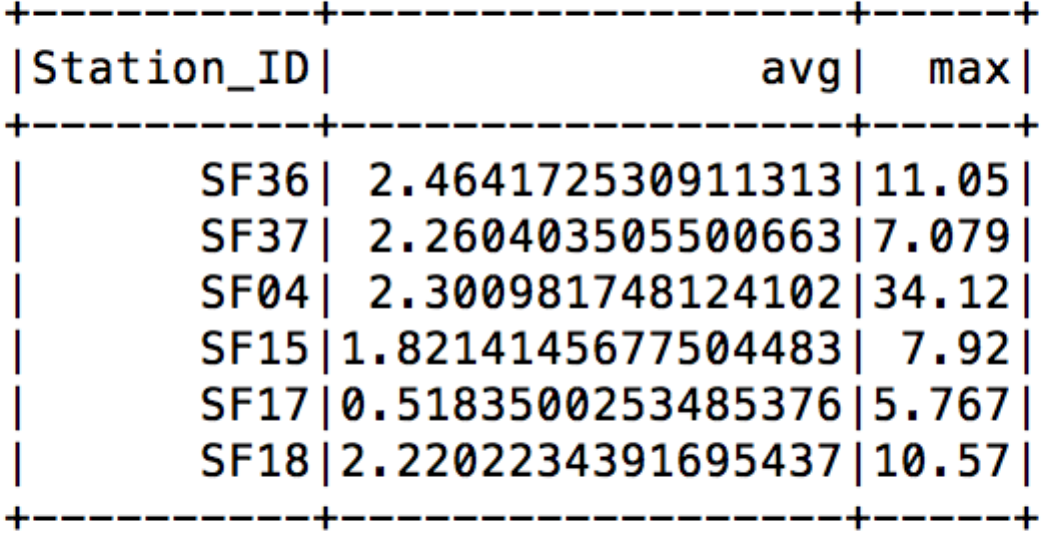
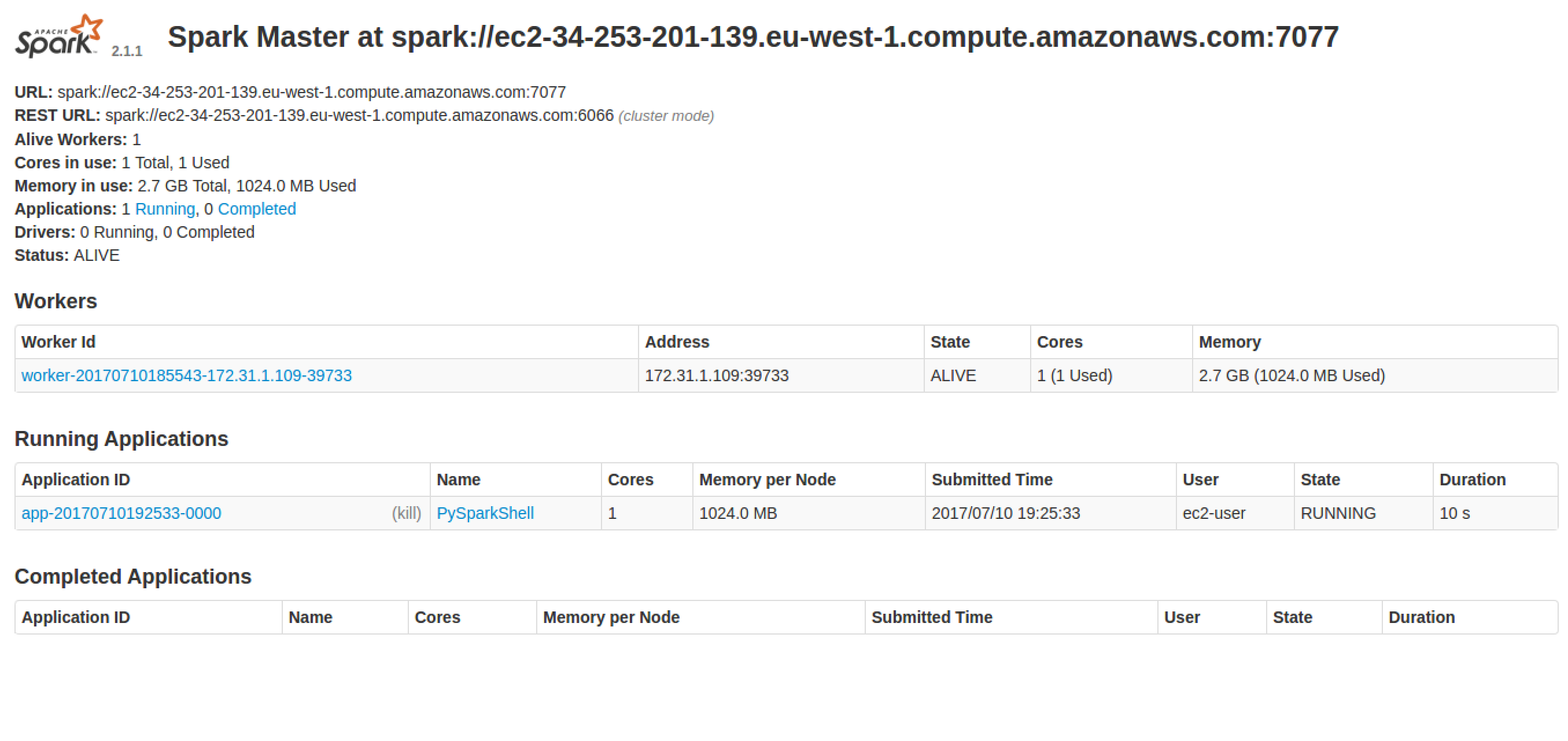
| SF15|Warnerville Switc...| Warnerville| 5| 2015-01-5? 00:20| 1.985| 6.895| 141.8| 0.5| 0.062|

+----------+--------------------+--------------+----------------+-----------------+---------------------+---------------------------+------------------+-------------------------+----------------------------+

only showing top 4 rows

1. We can also convert the DataFrame into an RDD, allowing us to do functional programming on it (map/reduce/etc)  
     
   winds = df.rdd
2. 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))
3. We can simply calculate the maximum values with this reducer:  
     
   maxes = mapped.reduceByKey(lambda a, b: a if (a>b) else b)
4. And once again collect / print:  
     
   for (k,v) in maxes.collect(): print k,v
5. You will see a bunch of log before the following appears:
6. You can also turn the response of a collect into a Python Map, which is handy. Try this:  
     
   maxes.collectAsMap()['SF04']
7. You can also try:  
   print maxes.collectAsMap()

**PART B – Using SQL**

1. There is an easier way to do all this if you are willing to write some SQL.
2. First we need to give our DataFrame a table name:  
   df.registerTempTable('wind')
3. Now we can use a simple SQL statement against our data.   
   ALL ON ONE Line type:  
     
   sqlContext.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()
4. Bingo you should see a lot of log followed by:  
   
5. Recap. So fat 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.
6. Find the IP address of the Spark Master. There are two ways. Firstly, it showed up in the console when you first launched the flintrock cluster:  
   [34.253.201.139] Configuring Spark master...  
      
   Alternatively, you can find it as “oxcloXX-sc-master” in the EC2 instances.
7. Go to <http://xx.xx.xx.xx:8080> using the master’s IP address.  
   You should see something like:
8. Quit the pyspark shell:   
   quit()
9. Exit the SSH session:  
   exit
10. 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-slave --num-slaves 1 oxcloXX-sc

To save you retyping all that spark code, look here: <https://freo.me/wind-sql>

1. If you are planning to do the **Jupyter on EC2 exercise** straight away, then you can start it now and use your existing flintrock/EC2 cluster. Otherwise please follow the next instruction to shut down the EC2 instances.
2. We must remember to stop our cluster as well (its costing money…)  
   From Ubuntu terminal where you started the Spark cluster  
     
   flintrock destroy oxcloXX-sc  
     
   Type y when prompted.
3. Congratulations, this lab is complete.