Exercise 6b

Getting Jupyter Notebooks to work with EC2

Prior Knowledge
Unix Command Line Shell
Exercise 6

Learning Objectives
Getting the benefits of Jupyter with Spark on EC2

Software Requirements
(see separate document for installation of these)

- EC2 credentials
- Flintrock
- Livy

Steps

1. Jupyter needs some extras to be able to talk to a remote Spark system. The Jupyter package is called **sparkmagic**

   Optionally, if you want to know more, the project is here: https://github.com/jupyter-incubator/sparkmagic

   Let's make sure that sparkmagic is up to date:

   ```
   sudo pip install --upgrade sparkmagic
   sudo jupyter nbextension enable --py --sys-prefix widgetsnbextension
   sudo pip install --upgrade numexpr
   ```

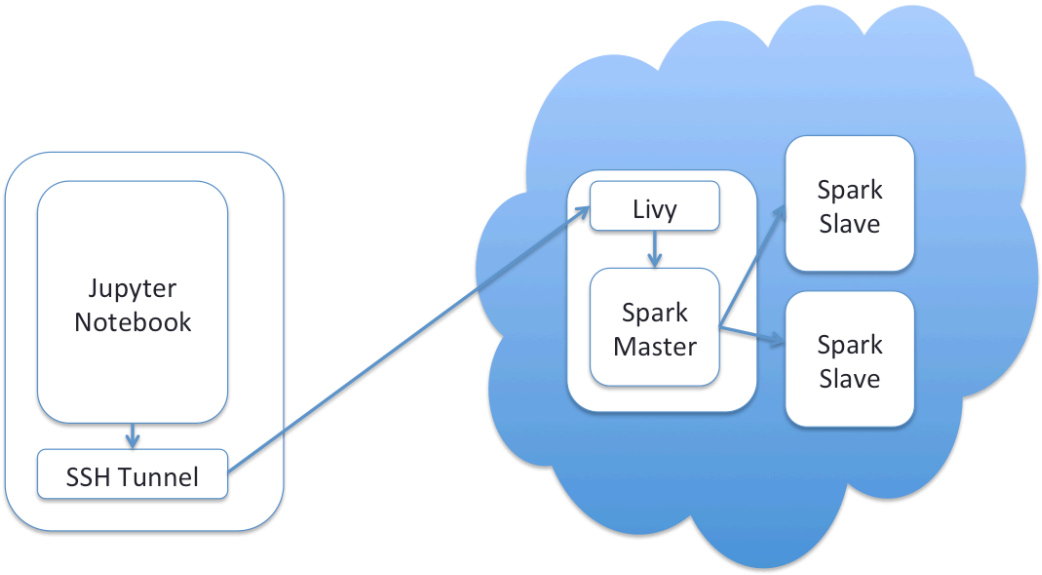
2. If you are still running the Flintrock EC2 cluster from Exercise 6, then skip the next step.

3. Launch a cluster

   ```
   cd ~/flintrock
   ./flintrock launch oxcloXX-sc
   ```

4. We are now going to install Livy, which is a RESTful interface to Spark. Livy will run on the Spark master node and listen on port 8998.

   The following diagram explains how this will work:



5. Find out the public IP of this instance from the EC2 console. Livy supports various security, so one option would be to enable security on Livy and then open up port 8998 by editing the security group. However, it is simpler and probably safer to use an *SSH Tunnel.*

   In a fresh Ubuntu terminal (i.e. not on your EC2 instance)**:**

   ```
   ssh -4  -i ~/keys/oxclo01.pem -L8998:localhost:8998 ec2-user@XX.XX.XX.XX
   ```

   (replace XX.XX.XX.XX with the public IP address of your Spark Master instance)

   This will give you a terminal window to control the Spark master (just like *flintrock login*), but in addition it will setup the SSH Tunnel.

   What this does is to take any traffic that comes to localhost:8998 and send it to the Livy server running on 8998 on the EC2 instance.

On the flintrock-logged in shell, type the following commands to install Livy and get it running (you can cut and paste from https://freo.me/oxclo-livy):

```
wget http://archive.cloudera.com/beta/livy/livy-server-0.3.0.zip
unzip livy-server-0.3.0.zip
cd livy-server-0.3.0
mkdir logs
echo "livy.spark.master = spark://0.0.0.0:7077" >> conf/livy.conf
bin/livy-server
```

The final part should result in something like:

```
17/07/14 06:59:49 INFO StateStore$: Using BlackholeStateStore for recovery.
17/07/14 06:59:49 INFO BatchSessionManager: Recovered 0 batch sessions. Next
session id: 0
17/07/14 06:59:49 INFO InteractiveSessionManager: Recovered 0 interactive
sessions. Next session id: 0
17/07/14 06:59:49 INFO InteractiveSessionManager: Heartbeat watchdog thread
started.
17/07/14 06:59:50 INFO WebServer: Starting server on http://ip-172-31-10-
147.eu-west-1.compute.internal:8998
```

If you had an EC2 setup you used regularly, you could configure this in userdata and have flintrock run it for you.

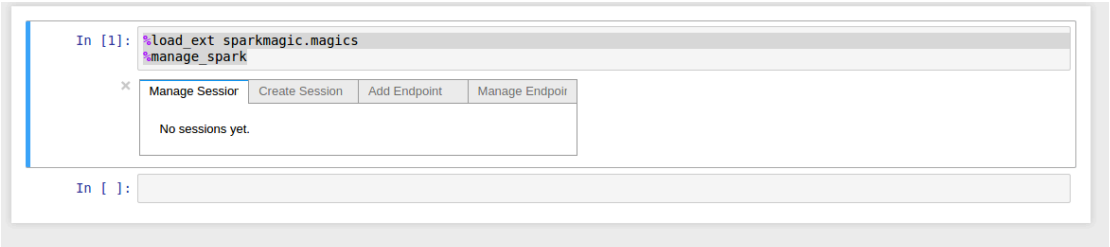
6. Now you can start Jupyter:

```
jupyter notebook
```

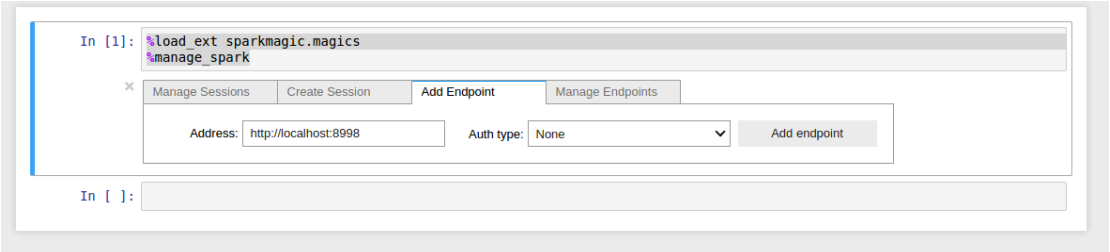
7. In the first cell, type:

```
%load_ext sparkmagic.magics
%manage_spark
```

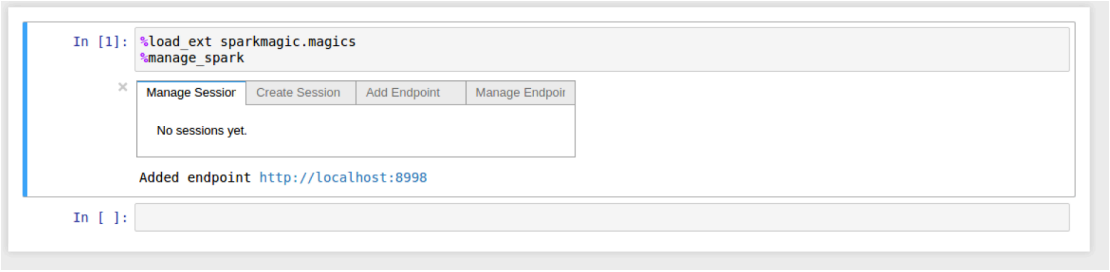
8. Run that cell.
You should see:



9. Click the **Add Endpoint** tab
Type **http://localhost:8998** as the endpoint :



10. Click the **Add Endpoint** button
    You should see:

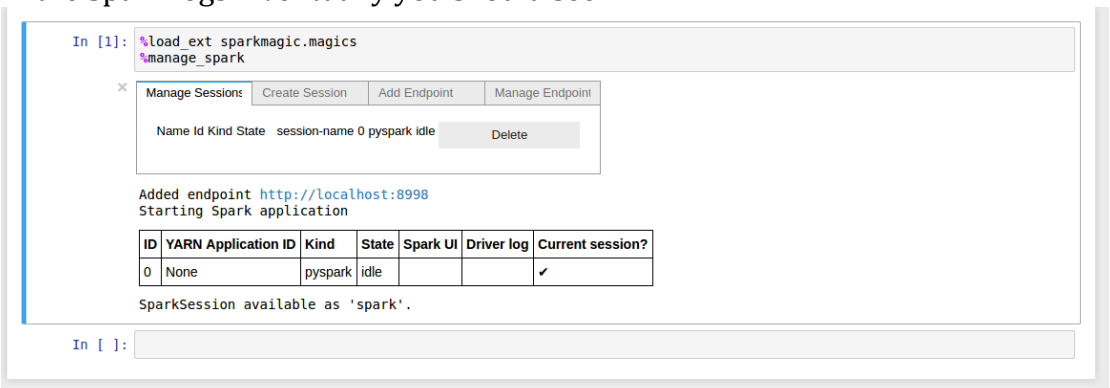


11. Now click the **Create Session** tab.
    Select the **Python** language.
    Use the following properties: (all one line):

```
{ "conf": {"spark.jars.packages": "com.amazonaws:aws-java-sdk-pom:1.10.34,org.apache.hadoop:hadoop-aws:2.7.2"}}
```

    This ensures that the Spark master can access S3 resources.

12. Click **Create Session** Button (you may have to scroll right). You will need to wait a bit. If you go to the Livy server window you can see the progress in the Spark logs. Eventually you should see:

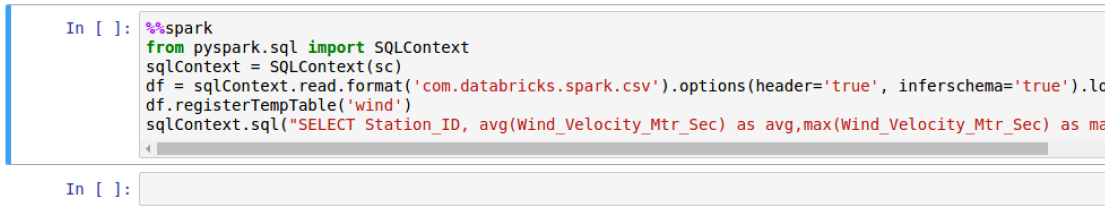


13. In the next cell type

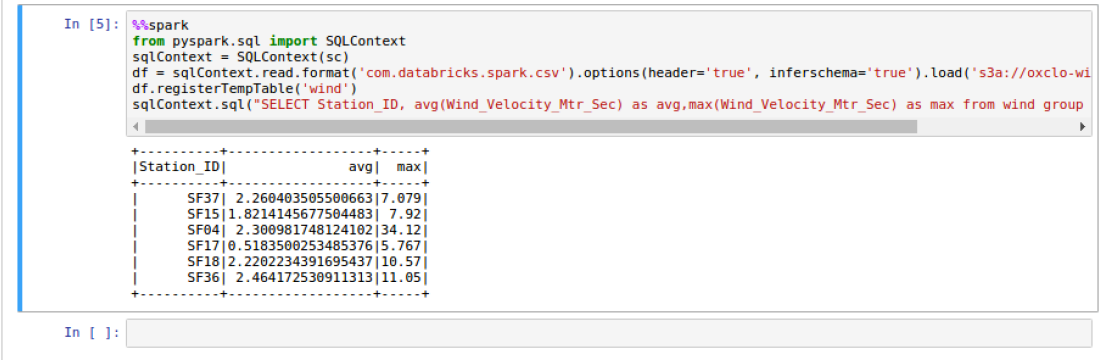
```
%%spark
```

    Then copy and paste the S3 wind analysis code from:
    https://freo.me/wind-sql

14. Your screen should look like:



15. Run that cell. You should see (after a while):



Extension

Take a look at the Sparkmagic instructions here to find out more secret commands. Try some of them out.
https://freo.me/sparkmagic

IMPORTANT

Don't forget to kill your cluster when done:

```
cd ~/flintrock
./flintrock destroy oxcloXX-sc
```