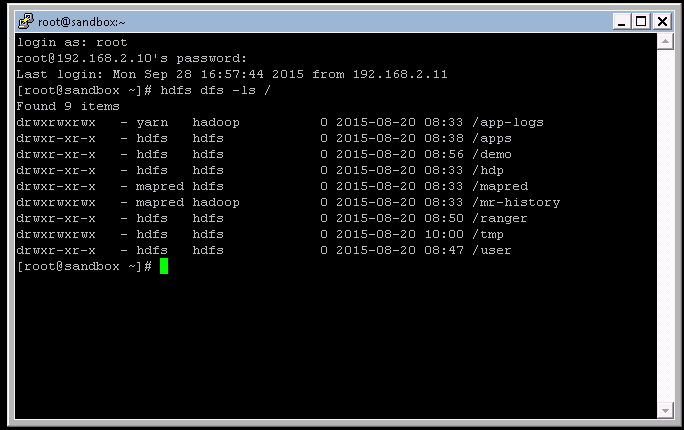
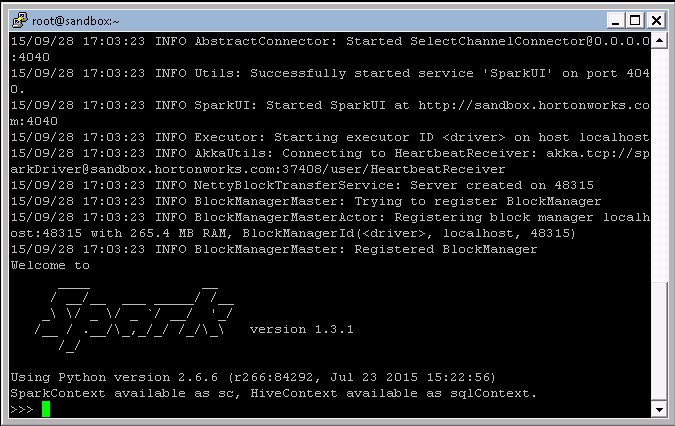
# Introduction to Spark and Machine Learning

## Environment Setup and Validation

1. Visit <https://labs.hds.com> in a web browser of your choice.
2. Click Specialty Labs under the Categories heading.
3. Look for the lab titled Hortonworks 2.3 with Spark.
4. Click the Start Lab button. Wait until configuration and setup dialogues complete. When initialization is complete you should see a Windows desktop.
5. Launch Putty by double clicking on the Putty icon on the Windows desktop.
6. Click on the session configuration titled Hadoop Server, then click on the Open button.
7. You will be prompted for credentials. Use root as the username and hadoop (lower case used here intentionally).
8. Ensure that Hadoop is properly installed and that HDFS is functioning normally by executing the following command and ensuring that what you see matches the screen shot below:
   1. hdfs dfs –ls /

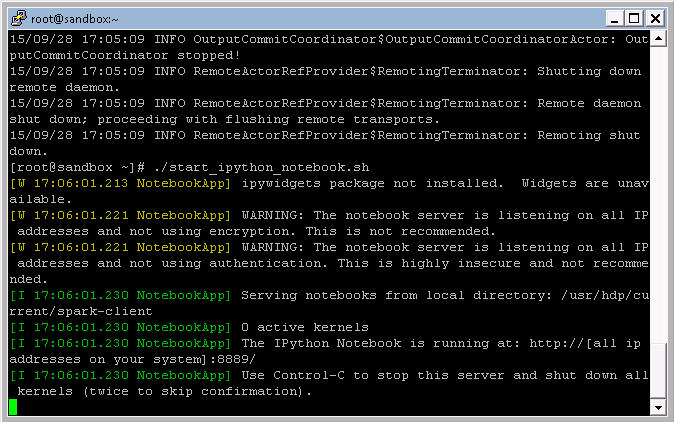


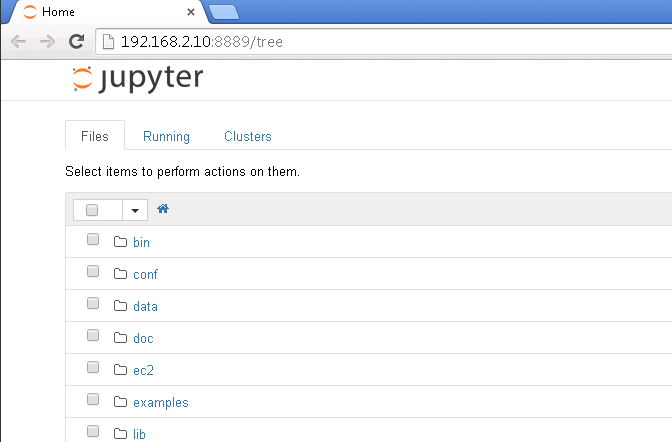
1. Ensure that Python and Spark are installed and operating correctly by executing the following command and ensuring that the what you see matches the screen shot below:
   1. pyspark



* 1. Type exit() to exit pyspark.

1. Finally, ensure that Jupyter, a web application for scientific computing that you’ll be using for the labs, is installed and is able to run. Execute the following command and ensure that what you see matches the screen shot below:
   1. ./start\_ipython\_notebook.sh



* 1. Start Chrome by double clicking on the Chrome icon on the Windows desktop. The Chrome homepage has been set to the local Jupyter instance. Ensure that Chrome shows content like:
  2. 

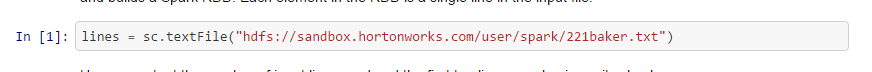
1. All validation steps are complete. You may now commence the lab exercises.

## Data Preparation

1. Download the following files to the VM desktop:
   1. <https://github.com/mikeydavison/sparkcourse/blob/master/221baker.txt>
   2. <https://github.com/mikeydavison/sparkcourse/blob/master/auto-mpg.data>
   3. <https://github.com/mikeydavison/sparkcourse/blob/master/iris.data>
2. Open Chrome, and go to the following URL to access the Hortonworks HUE UI
   1. <http://192.168.2.10:8000>
3. Click on the File Browser icon (5th from the left, near a pig and HCAT icon, a brown filing cabinet) to open the HUE file browser UI.
4. Navigate to the spark user home directory in HDFS by first clicking the green user hyperlink to navigate up one directory level. Then click on the spark hyperlink
5. Click on the Upload button at the top right of the UI. When prompted to select files, browse to the VM desktop and select the files downloaded in steps 1a-1c.
6. Connect to the Linux guest following the instructions in Environment Setup and Validation
7. Execute the following command to ensure that the data files are in the proper location:
   1. hdfs dfs –ls /user/spark
   2. You should see 221Baker.txt, auto-mpg.data, and iris.data listed.

## Lab 1 – Word Count in Spark

1. Open Google Chrome. Ensure that the Jupyter home page is rendered. If Chrome does not show the Jupyter home page, ensure that step 10 from Environment Setup and Validation has been successfully completed.
2. Download the lab 1 ipython notebook to the VM desktop
   1. <https://github.com/mikeydavison/sparkcourse/blob/master/spark_wc_lab1.ipynb>
3. In Chrome on the Jupyter home page, click the Upload button on the top right of the screen. Select the ipython notebook from step 2.
4. Once the upload completes, click the file named spark\_wc\_lab1.ipynb in the Jupyter file view.
5. Click on the first grey code box that appears as follows



1. Click the play button (sideways black triangle) towards the top of the page. Doing so will execute the python code in the grey box. The number in brackets [1] to the left of the code will briefly change to a [\*] until the code finishes executing.
2. Once the execution completes, click on each successive grey code box, then play, to execute the lab. Take care at each step to read the preceding instructions in order to understand what the code is doing.

## Lab 2 – Regression and Classification in Spark

1. Open Google Chrome. Ensure that the Jupyter home page is rendered. If Chrome does not show the Jupyter home page, ensure that step 10 from Environment Setup and Validation has been successfully completed.
2. Download the lab 2 ipython notebook to the VM desktop
   1. <https://github.com/mikeydavison/sparkcourse/blob/master/spark_ml_lab2.ipynb>
3. In Chrome on the Jupyter home page, click the Upload button on the top right of the screen. Select the ipython notebook from step 2.
4. Once the upload completes, click the file named spark\_ml\_lab2.ipynb in the Jupyter file view.
5. Click on the first grey code box that appears as follows



1. Click the play button (sideways black triangle) towards the top of the page. Doing so will execute the python code in the grey box. The number in brackets [1] to the left of the code will briefly change to a [\*] until the code finishes executing.
2. Once the execution completes, click on each successive grey code box, then play, to execute the lab. Take care at each step to read the preceding instructions in order to understand what the code is doing.