**Clairvoyant**

**Intro to Apache Spark Workshop:**

**Exercises**

# General Notes

The Intro to Apache Spark workshop uses the VM CDH5.0 Quick Start (Cloudera’s Distribution, including Apache Hadoop 5.X) installed in Psuedo-Distributed mode. The VM comes pre-installed with Apache Spark running on Apache Hadoop.

**Instructions to Install the VM**

1. Download VirtualBox from: <https://www.virtualbox.org/wiki/Downloads>
   1. Pick the appropriate binary for your operating system.
   2. Then follow the prompts for installing VirtualBox.
2. Once the installation is complete, download the CDH5.0 VM from the following URL:

<http://www.cloudera.com/content/cloudera/en/documentation/core/latest/topics/cloudera_quickstart_vm.html>

You may need to download 7-zip from [www.7-zip.org](http://www.7-zip.org) to extract the Cloudera VM. After uncompressing, you will get a file called “cloudera-quickstart-vm-5.X.X-0-virtualbox.ovf.” Move it to location of your choice in File System.

1. Once you have extracted the VM, we will load the VM into VirtualBox.
   1. Open Virtual Box and click on **File -> Import Appliance...**
   2. From the file dialog open **cloudera-quickstart-vm-5.X.X-0-virtualbox.ovf** located in the decompressed (or unzipped) Cloudera VM download.
2. Setup "Network Adapter 2" in Network Settings in the Virtual Box as "Host-only Adapter." Another option is bridged but it has a bug in Mac when using a wireless connection.
   1. If when setting up the "Host-only Adapter," the "Name" drop down is showing only "Not selected", cancel and go back to Virtual Box preferences ("Virtual Box -> preferences ->network).

* 1. Select **Host-Only Networks** then **Add** and a new entry will be created (something like "vboxnet0").
  2. Click **OK**. Now go back to Network Settings on the VM. This time Adapter 2 should show vboxnet0 in the "Name" drop down box. Select "vboxnet0."

1. Update Virtual Box to the latest version if you are not able to add as described in step 4. The menu might be present in an older version, but may throw an exception/error message when you attempt to add.
2. Add **Port 50010** to NAT adapter.
3. Open the VM Network Settings.
   1. Go to **Adapter 1**, 'Attached to:' NAT should display.
   2. Add **Port 50010** to Port Forwarding.
4. Start the VM.
5. Once the VM starts up, you should see the Desktop within VirtualBox. This is your sandbox to play with Hadoop.
6. Test to ensure the VM was setup successfully. Open Terminal and run the following commands:
   1. $ hadoop version
      1. If the setup is successful, it will print the current version.
   2. $ spark-shell
      1. In the shell that opens run command:

scala> **sc**

* + - 1. If the setup is successful, it will print a SparkContext object

# Exercise 1 – Access Logs

Analyze the access.log file and calculate the following:

* Count how many times the “/health” URL was hit
* Get all events that occurred on May 19th 2014 and save them to HDFS

Access log file can be found in two locations:

* In the spark-workshop-data.zip file provided, in the “logs” subdirectory
* In HDFS (on the VM provided) at

/user/cloudera/spark-workshop-data/logs/access.log

# Exercise 2 – Joining Datasets

Using the README.md and CHANGES.txt, find out how many time the word “Spark” shows up in both of the files together by following the bellow steps:

1. Create RDD’s to filter each file for the keyword “Spark”
2. Perform a WordCount on each of the resulting datasets so the results are (K, V) pairs of (word, count)
3. Join the two RDDs

Files can be found in two locations:

* In the spark-workshop-data.zip file provided, in the “spark” subdirectory
* In HDFS (on the VM provided) at

/user/cloudera/spark-workshop-data/spark/

# Exercise 3 – Shared Variables

Provide an exercise for participants that includes using broadcast variables and accumulators

# Exercise 4 – Spark Submit

Provide an exercise for participants that includes putting everything together