Hadoop Virtual Machine Guide

CSE 4/587

Prepared by Jacob Condello

VM SETUP

1. Download and install VirtualBox [1]
2. Download the virtual machine from UBbox [2]
3. Import Hadoop.ova by either double clicking it or from VirtualBox selecting File->Import Appliance
4. After setup run the virtual machine named "Hadoop"
5. Login with user "hadoop" password "hadoop"

HADOOP SETUP

1. Make sure you have access to a text editor. The VM comes with mousepad or you can install a different one.
2. Add the following lines to the end of your ~/.bashrc file:
   ```
   export JAVA_HOME=/usr/lib/jvm/java-6-openjdk-i386
   export HADOOP_CLASSPATH=$JAVA_HOME/lib/tools.jar
   ```
3. Apply changes by running:
   ```
   source ~/.bashrc
   ```
4. Start Hadoop by running:
   ```
   start-hadoop.sh
   ```
5. Make a new directory on hdfs for the input files:
   ```
   hdfs dfs –mkdir –p ~/input/
   ```
6. Copy the local directory books/ to the hdfs directory input/:
   ```
   hdfs dfs –put ~/books/ ~/input
   ```
7. Make sure you copied them to the right place:
   ```
   hdfs dfs –ls ~/input
   ```

MR CODE

1. Create a file named WordCount.java and copy in the source code for Example: WordCount v1.0 from hadoop.achache.org [3]
2. Run the following to compile WordCount.java:
   ```
   hadoop com.sun.tools.javac.Main WordCount.java
   ```
   Note: If you get an error at this point, it is likely either your java version, JAVA_HOME, or HADOOP_CLASSPATH that aren't configured correctly.

3. Run the following to create a jar:
```
jar cf wc.jar WordCount*.class
```
4. You can now run your MR job with the following:
```
hadoop jar wc.jar WordCount ~/input/books ~/output1
```
Notice that each time you run a MR job you need to either specify a new output file or delete the old one
5. You can now view the results by running:
```
hdfs dfs –cat ~/output1/*
```
Or if you need the results locally, copy from hdfs then cat:
```
hdfs dfs –get ~/output1
cat part-r-00000
```

REFERENCES

[1] https://www.virtualbox.org/wiki/Downloads

[2] https://buffalo.app.box.com/s/t72vy6py6r6v852anwa55whl4mmibfaw

[3] https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html