SETTING UP A SPARK CLUSTER:

Assuming you already have a Hadoop cluster set up, setting up Spark to run on top of it is fairly straight-forward.
These were the steps I took to install Spark on the CS machines on campus

1) Download Spark from http://spark.apache.org/downloads.html (I'm using 1.6.1, but any version should work)

2) Export the location of the Spark directory as a shell variable (i.e, edit your bashrc file and add the line "export SPARK_HOME=<path to Spark installation>"

3) Within SPARK_HOME, navigate into conf and open "spark-defaults.template". You will need to edit the following lines by uncommenting them and changing the values:

| | |
|---|---|
| spark.master | spark://HDFSNamenode:anyPortNum |
| spark.eventLog.enabled | true |
| spark.eventLog.dir | hdfs://HDFSNamenode:NamenodeListeningPort/history |
| spark.serializer | org.apache.spark.serializer.KryoSerializer |
| spark.executor.memory | 4g |

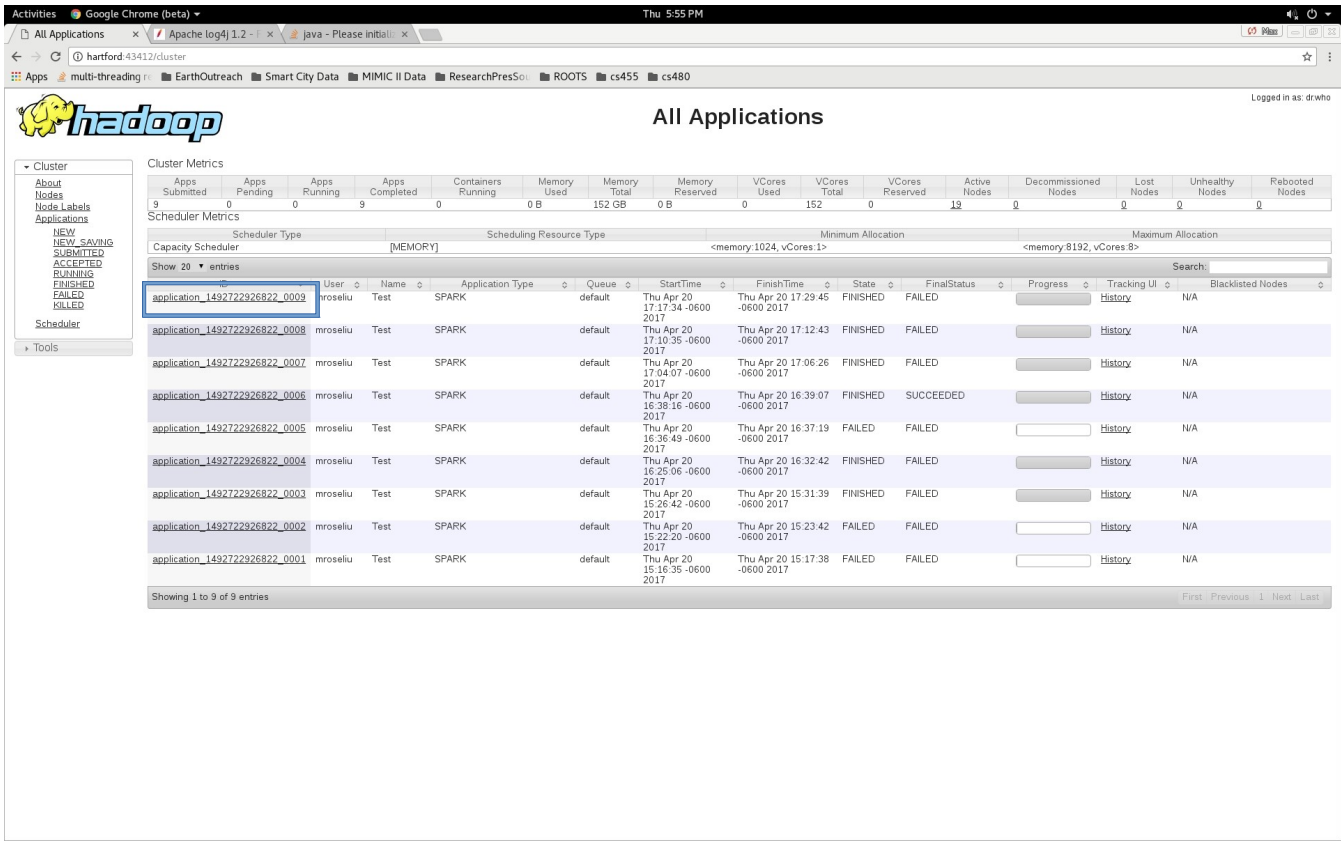
4) Save "spark-defaults.template" as "spark-defaults.conf"

5) Copy your "slaves" file from your hadoop configuration into the conf directory of SPARK_HOME.

6) Now you should be ready to launch your cluster.
    6a) First start up HDFS ($HADOOP_HOME/sbin/start-dfs.sh)
    6b) Start Yarn ($HADOOP_HOME/sbin/start-yarn.sh)
    6c) Start Spark ($SPARK_HOME/sbin/start-all.sh)

Hopefully there are no errors at this point. If there are, consult log files.

7) Check that Spark is up and running by going to your Yarn manager's web portal (the port number can be found in yarn-site.xml, property name: yarn.resourcemanager.webapp.address). It should look something like this:




You can view the status of any job by clicking the application number. You can then see the logs associate with that task to resolve errors.

8) To submit a job, create your class and also create a jar file. Make sure that HDFS, Yarn, and Spark are all up and running. The command that I have been using is:

$SPARK_HOME/bin/spark-submit --class <name of your main class> --master yarn --deploy-mode cluster <path to jar> <arguments to your main program>