Documentation for setting up the cluster

Install Java:

On each node:

```
sudo apt-get update
java -version
sudo apt-get install default-jdk
sudo update-alternatives --config java
(copy: /usr/lib/jvm/java-7-openjdk-amd64)
emacs .bashrc
JAVA_HOME="YOUR_PATH"
source .bachrc
java -version
```

Install SBT:

On each node:

```
echo "deb https://dl.bintray.com/sbt/debian /" | sudo tee -a /etc/apt/
sources.list.d/sbt.list
sudo apt-key adv --keyserver hkp://keyserver.ubuntu.com:80 --recv 642AC823
sudo apt-get update
sudo apt-get install sbt
```

Install Scala:

On each node:

```
wget www.scala-lang.org/files/archive/scala-2.11.7.deb
sudo dpkg -i scala-2.11.7.deb
```

Install Hadoop:

On each node:

```
wget http://apache.mirrors.tds.net/hadoop/common/hadoop-2.7.1/
```

```
hadoop-2.7.1.tar.gz
sudo tar zxvf hadoop-2.7.1.tar.gz
sudo emacs .bashrc
(.bashrc should have following line:)
        export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
        export PATH=$PATH:$JAVA_HOME/bin

        export HADOOP_HOME=/home/sgangarapu/hadoop-2.7.1
        export PATH=$PATH:$HADOOP_HOME/bin
        export HADOOP_CONF_DIR=/home/sgangarapu/hadoop-2.7.1/etc/hadoop

source .bashrc
```

Hadoop Configurations:

Here are the following files to focus on:

```
$HADOOP_CONF_DIR/hadoop-env.sh
$HADOOP_CONF_DIR/core-site.xml
$HADOOP_CONF_DIR/yarn-site.xml
$HADOOP_CONF_DIR/mapred-site.xml
```

Common Hadoop Configurations on all Nodes:

On each node:

```
sudo emacs $HADOOP_CONF_DIR/hadoop-env.sh
        export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64

sudo emacs $HADOOP_CONF_DIR/core-site.xml
        <configuration>
           <property>
              <name>fs.defaultFS</name>
              <value>hdfs://spark.rcg.usm.maine.edu:9000</value>
           </property>
        </configuration>
```

```
sudo emacs  $HADOOP_CONF_DIR/yarn-site.xml
        <configuration>

        <!– Site specific YARN configuration properties –>

          <property>
            <name>yarn.nodemanager.aux-services</name>
            <value>mapreduce_shuffle</value>
          </property>
          <property>
            <name>yarn.nodemanager.aux-
              services.mapreduce.shuffle.class</name>
             <value>org.apache.hadoop.mapred.ShuffleHandler</value>
           </property>
          <property>
            <name>yarn.resourcemanager.hostname</name>
            <value>spark.rcg.usm.maine.edu</value>
           </property>
        </configuration>

sudo cp $HADOOP_CONF_DIR/mapred-site.xml.template
        $HADOOP_CONF_DIR/mapred-site.xml

sudo emacs $HADOOP_CONF_DIR/mapred-site.xml
        <configuration>
          <property>
            <name>mapreduce.jobtracker.address</name>
            <value>spark.rcg.usm.maine.edu:54311</value>
          </property>
          <property>
            <name>mapreduce.framework.name</name>
            <value>yarn</value>
          </property>
        </configuration>
```

NameNode specific configurations:

```
sudo cat /etc/hosts

        127.0.0.1 localhost
        spark.rcg.usm.maine.edu spark
        172.20.132.2    workernode0
        172.20.132.3    workernode1
        172.20.132.4    workernode2

sudo emacs $HADOOP_CONF_DIR/hdfs-site.xml
        <configuration>
          <property>
            <name>dfs.replication</name>
            <value>3</value>
          </property>
          <property>
            <name>dfs.namenode.name.dir</name>
            <value>file:///usr/local/hadoop/hadoop_data/hdfs/
              namenode</value>
          </property>
        </configuration>

sudo mkdir -p $HADOOP_HOME/hadoop_data/hdfs/namenode
sudo touch $HADOOP_CONF_DIR/masters
sudo emacs $HADOOP_CONF_DIR/masters
        spark
sudo emacs $HADOOP_CONF_DIR/slaves
        workernode0
        workernode1
        workernode2
sudo chown -R sgangarapu $HADOOP_HOME
```

DataNode Specific Configurations

On each worker node:

```
sudo emacs $HADOOP_CONF_DIR/hdfs-site.xml
        <configuration>
          <property>
```

```
    <name>dfs.replication</name>
    <value>3</value>
  </property>
  <property>
    <name>dfs.datanode.data.dir</name>
    <value>file:///usr/local/hadoop/hadoop_data/hdfs/
      datanode</value>
  </property>
</configuration>
```

```
sudo mkdir -p $HADOOP_HOME/hadoop_data/hdfs/datanode
sudo chown -R sgangarapu $HADOOP_HOME
```

Start Hadoop Cluster:

On NameNode:

```
hdfs namenode -format
$HADOOP_HOME/sbin/start-dfs.sh
```

Namenode UI - http://spark.rcg.usm.maine.edu:50070/

Install Spark:

On each node:

```
wget http://apache.mirrors.tds.net/spark/spark-1.4.1/spark-1.4.1-bin-hadoop2.4.tgz
sudo emacs .bashrc
        export SPARK_HOME=/home/sgangarapu/spark-1.4.1-bin-hadoop2.4
        export PATH=$PATH:$SPARK_HOME/bin
source .bashrc
sudo chown -R sgangarapu $SPARK_HOME
```

Spark Configurations:

Common Hadoop Configurations on all Nodes:

```
sudo cp $SPARK_HOME/conf/spark-env.sh.template $SPARK_HOME/
```

```
        conf/spark-env.sh
sudo emacs $SPARK_HOME/conf/spark-env.sh
        export JAVA_HOME=/usr/lib/jvm/java-7-openjdk-amd64
        export SPARK_PUBLIC_DNS="current_node_public_dns"
        export SPARK_WORKER_CORES=6
```

Master specific configurations:

```
touch $SPARK_HOME/conf/slaves
sudo emacs $SPARK_HOME/conf/slaves
        172.20.132.2
        172.20.132.3
        172.20.132.4
```

Start Spark Cluster:

On Master:

```
$SPARK_HOME/sbin/start-master.sh
```

On Workers:

```
$SPARK_HOME/sbin/start-slave.sh spark://spark.rcg.usm.maine.edu:7077
```

Spark Master UI - http://spark.rcg.usm.maine.edu:8080/

Setup Jupyter Server on Spark Cluster:

On each node:

Install Python:

```
sudo apt-get install python-dev python-pycurl python-simplejson python-pip
        libzmq-dev
sudo apt-get purge libzmq-dev
sudo -H pip install tornado pyzmq
sudo pip install “ipython[all]”
sudo -H pip install jinja2
sudo apt-get install libfreetype6-dev libxft-dev
sudo apt-get install git
```

Install ZeroMQ:

```
        sudo apt-get install libtool autoconf automake uuid-dev build-
            essential
        wget http://download.zeromq.org/zeromq-3.2.2.tar.gz
        tar zxvf zeromq-3.2.2.tar.gz && cd zeromq-3.2.2
        ./configure
        sudo make && make install

    sudo apt-get install python-pip
    sudo -H pip install path.py
    sudo -H pip install -U setuptools
    sudo -H pip install matplotlib jsonschema
    sudo -H pip install scikit-learn
    sudo -H pip install numpy scipy pandas
    sudo -H pip install path.py

    sudo -H pip install jupyter
```

On Master node:

```
sudo emacs .bashrc
    export PYSPARK_SUBMIT_ARGS='--master spark://
        spark.rcg.usm.maine.edu:7077 pyspark-shell'
source .bashrc
ipython profile create pyspark        (create python profile)
sudo emacs $HADOOP_HOME/.ipython/profile_pyspark/
  ipython_notebook_config.py
    c.NotebookApp.ip = "spark.rcg.usm.maine.edu"
    c.NotebookApp.open_browser =False
    c.NotebookApp.port = 42424
sudo touch $HADOOP_HOME/.ipython/profile_pyspark/startup/00-pyspark-
  setup.py
sudo emacs $HADOOP_HOME/.ipython/profile_pyspark/startup/00-pyspark-
  setup.py
    import os
    import sys

    spark_home = os.environ.get('SPARK_HOME)
```

```
if not spark_home:
    raise ValueError('SPARK_HOME environment variable is not set')
    sys.path.insert(0, os.path.join(spark_home, 'python'))
    sys.path.insert(0, os.path.join(spark_home,'python/lib/py4j-0.8.2.1-
      src.zipexecfile(os.path.join(spark_home,'python/pyspark/shell.py
```

Launch Jupyter Notebook:

```
jupyter notebook
```

Command for spark-csv to work on jupyter:

```
PYSPARK_DRIVER_PYTHON=ipython
PYSPARK_DRIVER_PYTHON_OPTS="notebook --no-browser —port=4242”
pyspark --packages com.databricks:spark-csv_2.10:1.1.0 --master spark://
spark.rcg.usm.maine.edu:7077
```