Spark Installation manual

Installing spark in a stand alone machine

Step 1: Check whether java is installed and JAVA_HOME and PATH is set

Step 2: Install Scala. Set SCALA_HOME and PATH in .\bashrc
You can download scala from the location,"http://www.scala-lang.org/download/"

Step 3: Verify the scala installation using the following command:
>scala -version

Step 4: Download Apache spark from the given location:
https://spark.apache.org/downloads.html
Choose a Spark release:2.2.0(Jul 11 2017)
Choose a package type:Pre-built for Apache hadoop 2.7 and later
Choose a download type: Direct Download
Download Spark: spark-2.2.0-bin-hadoop2.7.tgz
[]

Step 5: Extract the Spark tar : sudo tar xvf spark-1.3.1-bin-hadoop2.6.tgz

Step 6: Move the spark files to local: sudo mv spark-1.3.1-bin-hadoop2.6 /usr/local/spark

Step 7: Set up the environment in ~/.bashrc
export PATH = $PATH:/usr/local/spark/bin

Step 8: Execute the bash file: source ~/.bashrc

Step 9: Verify the spark installation: spark-shell

Installing spark in a cluster

Configuring the master node

Step 1: Check if java,scala and spark,ssh are installed

Step 2: Edit the hosts file /etc/hosts
>sudo nano /etc/hosts
Now add entries of master and slaves
<MASTER-IP> master
<SLAVE01-IP> slave01
<SLAVE02-IP> slave02

(NOTE: In place of MASTER-IP, SLAVE01-IP, SLAVE02-IP put the value of the corresponding IP)

Step 2: Edit the conf/spark-env.sh file
Create a copy of template of spark-env.sh and rename it:
>cp spark-env.sh.template spark-env.sh
Set the JAVA_HOME in spark-env.sh file
>export JAVA_HOME=<path-of-Java-installation> (eg: /usr/lib/jvm/java-7-oracle/)
export SPARK_WORKER_CORES=8

Step 3: Add Slaves
Create configuration file slaves (in $SPARK_HOME/conf/) and add following entries:
slave01

slave02

Copy setups from master to all slaves

Run these commands on master

Create tarball of configured setup

$ tar czf spark.tar.gz spark-2.0.0-bin-hadoop2.6

Copy the configured tarball on all the slaves

$ scp spark.tar.gz slave01:~

(Note: Mention the IP of slave1 and slave2 in place of slave01 and slave02)

Install Spark On Slaves

Step 1: Add entries in hosts file

Step 2: Install Java

Step 3: Install Scala

Step 4: Un-tar configured spark setup on all the slaves

$tar xzf spark.tar.gz

Cluster has been configured successfully.

Start Spark Cluster

Step 1: Start Spark Services in the master

>sbin/start-all.sh

Step 2: Browse the Spark UI to know about worker nodes, running application, cluster resources.

http://MASTER-IP:8080/

Step 3: Spark application UI

http://MASTER-IP:4040/