**Hadoop & Spark Installation**

Master

1. **Hadoop**

10.208.28.86

* 1. **Prerequisites**

1. Download Hadoop and spark binaries

Dnode

<https://hadoop.apache.org/release/2.7.7.html>

<https://archive.apache.org/dist/spark/spark-2.0.1/>

10.208.28.89

**$tar zxvf spark-2.1.0-bin-hadoop2.7.tgz**

**$tar zxvf hadoop-2.7.7.tar.gz**

1. Hosts Files
   1. Add below to /etc/hosts files in both machines

**10.208.28.86 master.hadoop.com**

**10.208.28.89 dnode2.hadoop.com**

1. Java JDK 8

**$java -version**

1. ssh password less between cluster nodes

**$ssh-copy-id root@10.208.28.89**

1. Service Configuration
   * 1. Environment Configs
        1. Bash Profile

**On 10.208.28.86 / 87**

**$vi .bash\_profile**

# JAVA env variables

export JAVA\_HOME=/usr/java/default/

export PATH=$PATH:$JAVA\_HOME/bin

export CLASSPATH=.:$JAVA\_HOME/jre/lib:$JAVA\_HOME/lib:$JAVA\_HOME/lib/tools.jar

## HADOOP env variables

export HADOOP\_HOME=/opt/hadoop

export HADOOP\_COMMON\_HOME=$HADOOP\_HOME

export HADOOP\_HDFS\_HOME=$HADOOP\_HOME

export HADOOP\_MAPRED\_HOME=$HADOOP\_HOME

export HADOOP\_YARN\_HOME=$HADOOP\_HOME

export HADOOP\_OPTS="-Djava.library.path=$HADOOP\_HOME/lib/native"

export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native

export HADOOP\_LOG\_DIR=/data/hadoop/logs/$USER

export PATH=$PATH:$HADOOP\_HOME/sbin:$HADOOP\_HOME/bin

**$source .bash\_profile**

* + - 1. **ssh password less [10.208.28.86]**

**$ssh-keygen -t rsa**

**$ssh-copy-id root@dnode2.hadoop.com**

* + 1. **HDFS configs on both machines**
       1. **/opt/hadoop/etc/hadoop/core-site.xml**

<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://master.hadoop.com:8020/</value>

</property>

</configuration>

* + - 1. **/opt/hadoop/etc/hadoop/hdfs-site.xml**

Master(10.208.28.86) Slave(10.208.28.89)

<configuration>

<property>

<name>dfs.data.dir</name>

<value>file:///data/hadoop/volume/datanode</value>

</property>

<property>

<name>dfs.name.dir</name>

<value>file:///data/hadoop/volume/namenode</value>

</property>

<property>

<name>dfs.replication</name>

<value>2</value>

</property>

</configuration>

<configuration>

<property>

<name>dfs.data.dir</name>

<value>file:///data/hadoop/volume/datanode</value>

</property>

<property>

<name>dfs.replication</name>

<value>2</value>

</property>

</configuration>

* + - 1. **/opt/hadoop/etc/hadoop/mapred-site.xml**

<?xml version="1.0"?>

<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>

<configuration>

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

</property>

</configuration>

* + 1. **Yarn Configs**
       1. **yarn-site.xml on both machines**

<configuration>

<property>

<name>yarn.resourcemanager.hostname</name>

<value>10.208.28.86</value>

</property>

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.resourcemanager.address</name>

<value>10.208.28.86:8032</value>

</property>

<property>

<name>yarn.resourcemanager.scheduler.address</name>

<value>10.208.28.86:8030</value>

</property>

<property>

<name>yarn.resourcemanager.resource-tracker.address</name>

<value>10.208.28.86:8031</value>

</property>

<property>

<name>yarn.log-aggregation-enable</name>

<value>true</value>

</property>

<property>

<name>yarn.log-aggregation.retain-seconds</name>

<value>172800</value>

</property>

<property>

<name>yarn.nodemanager.log.retain-seconds</name>

<value>259200</value>

</property>

<property>

<name>yarn.nodemanager.log-aggregation.roll-monitoring-interval-seconds</name>

<value>60</value>

</property>

<property>

<name>yarn.nodemanager.resource.memory-mb</name>

<value>8200</value>

</property>

<property>

<name>yarn.scheduler.maximum-allocation-mb</name>

<value>6100</value>

</property>

<property>

<name>yarn.scheduler.maximum-allocation-vcores</name>

<value>6</value>

</property>

<property>

<name>yarn.nodemanager.pmem-check-enabled</name>

<value>false</value>

</property>

<property>

<name>yarn.nodemanager.vmem-check-enabled</name>

<value>false</value>

</property>

<property>

<name>yarn.nodemanager.log-aggregation.compression-type</name>

<value>gz</value>

</property>

<property>

<name>yarn.nodemanager.log-dirs</name>

<value>/data/hadoop/yarn/log</value>

</property>

</configuration>

* + - 1. **Slaves Files**

**#Master(10.208.28.86)**

**$ vi /opt/hadoop/etc/hadoop/slaves**

**master.hadoop.com**

**dnode2.hadoop.com**

* + - 1. **-Edit Hadoop Environment File**

**vi /opt/hadoop/etc/hadoop/hadoop-env.sh**

**export JAVA\_HOME=/usr/java/default**

* 1. **Launch**

**#Master(10.208.28.86)**

**$hdfs namenode -format**

**$start-dfs.sh**

**$start-yarn.sh**

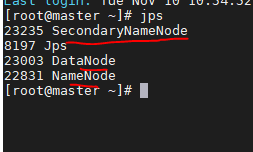
* 1. **Checkup services**

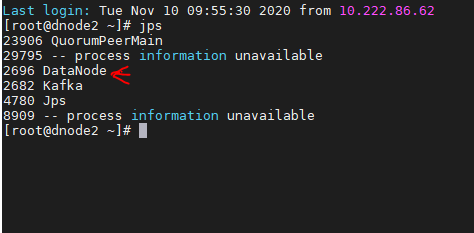
**#dnode1(10.208.28.89)**

**$jps**

**#Master(10.208.28.86)**

**$jps**





**open that on browser:**

http://10.208.28.86:50070

http://10.208.28.86:8088

http://10.208.28.86:8042

1. **Spark (all steps applied on all cluster nodes[86/89])**
   1. set spark and hadoop ENV integration configs

$vi .bash\_profile

export SPARK\_HOME=/opt/spark

export PATH=$SPARK\_HOME/bin:$PATH

export HADOOP\_CONF\_DIR=/opt/hadoop/etc/hadoop

export LD\_LIBRARY\_PATH=/opt/hadoop/lib/native:$LD\_LIBRARY\_PATH

export PATH=$PATH:$SPARK\_HOME/bin

$source .bash\_profile

* 1. spark configs

$cp $SPARK\_HOME/conf/spark-defaults.conf.template $SPARK\_HOME/conf/spark-defaults.conf

$vi $SPARK\_HOME/conf/spark-defaults.conf

$hdfs dfs -mkdir /spark2-history

--------------------------------

spark.driver.memory 512m

spark.yarn.am.memory 512m

spark.executor.memory 512m

spark.eventLog.dir hdfs://master.hadoop.com:8020/spark2-history

spark.eventLog.enabled true

spark.history.fs.logDirectory hdfs://master.hadoop.com:8020/spark2-history

spark.history.fs.update.interval 10s

spark.history.ui.port 18080

spark.history.fs.cleaner.enabled true

spark.history.fs.cleaner.maxAge 1h

spark.history.fs.cleaner.interval 1h

--------------------------------

* 1. Edit Spark Environment

$ mkdir -p /opt/tmp/spark

$cp $SPARK\_HOME/conf/spark-env.sh.template $SPARK\_HOME/conf/spark-env.sh

$vi $SPARK\_HOME/conf/spark-env.sh

export SPARK\_LOCAL\_DIRS=/opt/tmp/spark

export HADOOP\_CONF\_DIR=/opt/hadoop/etc/hadoop

export YARN\_CONF\_DIR=/opt/hadoop/etc/hadoop

* 1. Environment Testing

**spark-submit --deploy-mode client --master yarn \**

**--class org.apache.spark.examples.SparkPi \**

**$SPARK\_HOME/examples/jars/spark-examples\_2.11-2.1.0.jar 10**