**HADOOP SINGLE NODE CLUSTER**

* Make sure that you have a right version of java installed on your machine.
* Disable IPv6

Folder Location: /etc/sysctl.conf

net.ipv6.conf.all.disable\_ipv6=1

net.ipv6.conf.default.disable\_ipv6=1

net.ipv6.conf.lo.disable\_ipv6=1

* Install ssh

sudo apt-get install ssh

* Enable ssh

sudo ssystemctl enable ssh

* start ssh

sudo ssystemctl start ssh

* To generate SSH Key

ssh-keygen -t rsa -p ""

* Download hadoop from hadoop.apache.org
* click on binary file and copy link of the file and download it
* Untar it
* Setup variables in ./.bashrc file

## JAVA\_HOME

export JAVA\_HOME=/usr/lib/jvm/java-8-openjdk-amd64

export PATH=$PATH:export PATH=$PATH:/usr/lib/jvm/java-8-openjdk-amd64/bin

## HADOOP\_HOME

export HADOOP\_HOME=/hadoop-2.7.3

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

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

export HADOOP\_CONF\_DIR=$HADOOP\_HOME/etc/hadoop

export HADOOP\_MAPRED\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_HOME=$HADOOP\_HOME

export HADOOP\_HDFS\_HOME=$HADOOP\_HOME

export YARN\_HOME=$HADOOP\_HOME

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

export HADOOP\_LOG\_DIR=$HADOOP\_HOME/logs

export HADOOP\_OPTS="-Djava.library.path=$HADOOP\_COMMON\_LIB\_NATIVE\_DIR"

* Configure property in core-site.xml
* Configure properties in hdfs-site.xml

<configuration>

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<property>

<name>dfs.permission</name>

<value>false</value>

</property>

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

<value>file:/hadoop-2.7.3/hadoopdata/hdfs/namenode</value>

</property>

<property>

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

<value>file:/hadoop-2.7.3/hadoopdata/hdfs/datanode</value>

</property>

</configuration>

* Configure mapred-site.xml
* Configure yarn-site.xml
* Format namenode
* I was getting an error “unable to create file in namenode”.Giving all the permission to hadoop folder resolved the issue

sudo chmod -R ugo+rw /hadoop-2.7.3

* start services

start-dfs.sh

start-yarn.sh

* To check if all nodes are active:

-Type command jps in terminal, that displays all the hadoop daemons which are running on the system.

ResourceManager, DataNode, Jps, SecondaryNamenodeNameNode,NodeManager

Execute your wordcount program

**AWS**

I was getting below error:

Exception in thread "main" java.lang.ClassNotFoundException: WordCount

at java.net.URLClassLoader.findClass(URLClassLoader.java:381)

at java.lang.ClassLoader.loadClass(ClassLoader.java:424)

at java.lang.ClassLoader.loadClass(ClassLoader.java:357)

at java.lang.Class.forName0(Native Method)

at java.lang.Class.forName(Class.java:348)

at org.apache.hadoop.util.RunJar.run(RunJar.java:214)

at org.apache.hadoop.util.RunJar.main(RunJar.java:136)

The issue was resolved after passing class name in the

EMR-->Steps-->Add Step-->Arguments-->Class name, Input Directory Path, Output Directory path

The major thing I learnt during the process are

1. Installation of single node hadoop cluster, only one Datanode running and setting up all the other nameNode.
2. How to run the program on the single node hadoop cluster.
3. Creation of security key pair
4. Running a Map/Reduce Job on AWS. Building a cluster on the cloud
5. The processing of data is faster in hadoop as the input data is on the local system.
6. Got brief idea about EC2, S3 and EMR.