**Running Hadoop Environment in EC2 instance**

Step – 1 : Launch an EC2 instance

A screenshot of a computer

Description automatically generated

Step – 2 : Choose ubuntu 22.04 ami

A screenshot of a computer

Description automatically generated

Step – 3 : Give the necessary security groups

A screenshot of a computer

Description automatically generated

Step – 4 : Launch the instance

A screenshot of a computer

Description automatically generated

Step – 5 : SSH into EC2 server

A screenshot of a computer

Description automatically generated

Step- 6 : Root user : sudo su –

apt update

apt install openjdk-8-jdk

Step – 7 : To set path

Nano ~/.bashrc

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

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

source ~/.bashrc

Step – 8 : Check java path

A screenshot of a computer program

Description automatically generated

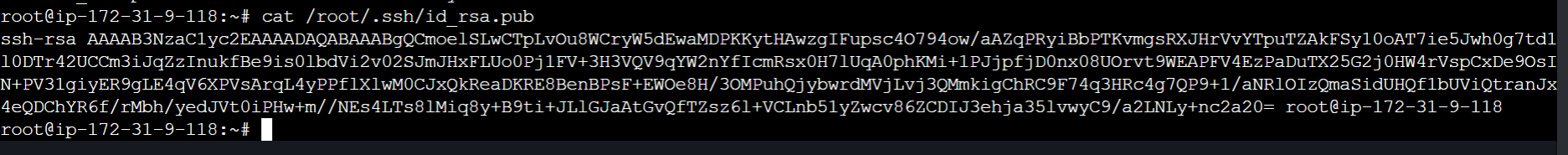
Step – 9 : Create pub key for the server

ssh-keygen -t rsa -P “”

A screenshot of a computer

Description automatically generated

Step – 10 : Now copy this id\_rsa.pub key to authorized\_keys



cat id\_rsa.pub >> authorized\_keys

Step – 11 : Now, ssh into server with

Ssh <public-ip> or ssh <private-ip>

A screenshot of a computer error

Description automatically generated

Step – 12 : Now download and install Hadoop

wget <https://dlcdn.apache.org/hadoop/common/hadoop-2.10.2/hadoop-2.10.2.tar.gz>

tar zxvf [https://dlcdn.apache.org/hadoop/common/hadoop-2.10.2/hadoop-2.10.2.tar.gz -C ./](https://dlcdn.apache.org/hadoop/common/hadoop-2.10.2/hadoop-2.10.2.tar.gz%20-C%20./)

A screenshot of a computer

Description automatically generated

Step – 13 : Now, save the Hadoop home environment variables in bashrc

nano ~/.bashrc

A screen shot of a computer

Description automatically generated

Step-14 : change java\_home in Hadoop-env.sh

A screen shot of a computer program

Description automatically generated

Step – 15 : Create a directory to save metadata when u run Hadoop commands and jobs

Mkdir -p Hadoop\_data/tmp

Step – 16 : Now, go to core-site.xml file and write the code. This will tell who is master

A screenshot of a computer

Description automatically generated

Step - 17 : create two directories for maintaining metadata for namenode and datanode

mkdir -p hadoop\_data/namenode

mkdir -p hadoop\_data/datanode

Step – 18 : Change the ownership permissions

Chown -R root:root /root

Step – 19 : modify hfds-site.xml file

A computer screen shot of a computer code

Description automatically generated

Step – 20 : Change the mapred-site.xml file

A screenshot of a computer

Description automatically generated

Step – 21 : Change yarn-site.xml -> To specify yarn node manager and yarn resource manager

A screenshot of a computer program

Description automatically generated

Step – 22 : Now the environment is read, start executing Hadoop cluster

hdfs namenode -format

This will format the hdfs

A computer screen with text on it

Description automatically generated

A screenshot of a computer screen

Description automatically generated

Step – 23 : Start the distributed file system by executing the cmd

Start-dfs.sh

This cmd will start namenode, secondary namenode and datanode

A screenshot of a computer

Description automatically generated

Step – 24 : now, start-yarn.xml cmd to start yarn process

This cmd will start resource manager and node manager

A screenshot of a computer program

Description automatically generated

Step – 25 : write jps command to check the currently running servers

A screenshot of a computer

Description automatically generated

Step – 26 : Now, copy the public IP of ur EC2 instance

A screenshot of a computer

Description automatically generated

Step – 27 : Paste the copied publicIP in the browser with :9870 at end

<public-ip>:9870 to access the hdfs webpage to check the datanodes,

A screenshot of a computer

Description automatically generated

Step – 28 : <public-ip>:8088 write this in the browser

This will get the following page

A screenshot of a computer

Description automatically generated

This page will tell what are the nodes running in the Hadoop cluster currently.

I successfully installed Hadoop environment inside the ec2 instance by installing dependencies like java and configuring files such as core-site.xml, hdfs-site.xml, yarn-site.xml and mapred-site.xml.

Now, we can perform Hadoop cmds like

Hdfs dfs -ls

Hdfs dfs -mkdir -p WordCount

Hdfs dfs -cp <local-path> <Hadoop-path> in site Hadoop environment.