Introduction to Big Data

ECAP456

Dr. Rajni Bhalla

Associate Professor

Learning Outcomes

After this lecture, you will be able to

- learn setup of the Hadoop Multi-Node cluster on a distributed environment
- learn how to create a system user account

Hadoop

- As the whole cluster cannot be demonstrated, we are explaining the Hadoop cluster environment using three systems (one master and two slaves); given below are their IP addresses.
  - Hadoop Master: 192.168.1.15 (hadoop-master)
  - Hadoop Slave: 192.168.1.16 (hadoop-slave-1)
  - Hadoop Slave: 192.168.1.17 (hadoop-slave-2)

Steps to have Hadoop Multi-Node cluster setup.

Installing Java

- Java is the main prerequisite for Hadoop. First of all, you should verify the existence of java in your system using

  “java -version”.

Creating User Account

- Create a system user account on both master and slave systems to use the Hadoop installation.

```
useradd Hadoop

passwd hadoop
```

Mapping the nodes

- You have to edit hosts file in /etc/ folder on all nodes, specify the IP address of each system followed by their host names.

Configuring Key Based Login

- Setup ssh in every node such that they can communicate with one another without any prompt for password.

```
su hadoop

ssh-keygen -t rsa

ssh-copy-id -i ~/.ssh/id_rsa.pub hadoop1@hadoop-master

ssh-copy-id -i ~/.ssh/id_rsa.pub hadoop_tp1@hadoop-slave-1

ssh-copy-id -i ~/.ssh/id_rsa.pub hadoop_tp2@hadoop-slave-2

chmod 0600 ~/.ssh/authorized_keys

exit
```

Steps

- Installing Hadoop
- Configuring Hadoop
  - Hadoop server must be configured
    - core-site.xml should be edited.
    - hdfs-site.xml file should be editted.
    - mapred-site.xml file should be editted.

Installing Hadoop on Slave Servers

- Install Hadoop on all the slave servers by following the given commands.

Configuring Hadoop on Master Server

- Open the master server and configure it by following the given commands.

- Configuring Master Node

Slave Node Configuration

```
vi etc/hadoop/slaves
```

```
hadoop-slave-1

hadoop-slave-2
```

Name Node format on Hadoop Master

Hadoop Services

- Starting Hadoop services on the Hadoop-Master.

```
cd $HADOOP_HOME/sbin

start-all.sh
```

Addition of a New DataNode in the Hadoop Cluster Networking

- Add new nodes to an existing Hadoop cluster with some suitable network configuration. suppose the following

- For New node Configuration:

IP address : 192.168.1.103

netmask : 255.255.255.0

hostname : slave3.in

Adding a User and SSH Access

- Add a User: “hadoop” user must be added and password of Hadoop user can be set to anything one wants.

```
useradd hadoop

passwd hadoop
```

To be executed on master

```
mkdir -p $HOME/.ssh
chmod 700 $HOME/.ssh
ssh-keygen -t rsa -P " -f $HOME/.ssh/id_rsa
cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
chmod 644 $HOME/.ssh/authorized_keys
```

Copy the public key to new slave node in hadoop user $HOME directory

```
scp $HOME/.ssh/id_rsa.pub hadoop@192.168.1.103:/home/hadoop/
```

To be executed on slaves

Login to hadoop. If not login to hadoop user

```
su hadoop ssh -X hadoop@192.168.1.103
```

Copy the content of public key into file "$HOME/.ssh/authorized_keys" and then change the permission for the same by executing the following commands

Set Hostname of New Node

- You can set hostname in file /etc/sysconfig/network
- On new slave3 machine
- NETWORKING = yes
- HOSTNAME = slave3.in
- To make the changes effective, either restart the machine or run hostname command to a new machine with the respective hostname (restart is a good option).

On slave3 node machine

- hostname slave3.in
- Update /etc/hosts on all machines of the cluster with the following lines –
- 192.168.1.102 slave3.in slave3
- Now try to ping the machine with hostnames to check whether it is resolving to IP or not.
- On new node machine –
- ping master.in

Start the DataNode on New Node

- Start the datanode daemon manually using $HADOOP_HOME/bin/hadoop-daemon.sh script. It will automatically contact the master (NameNode) and join the cluster. We should also add the new node to the conf/slaves file in the master server. The script-based commands will recognize the new node.

Start the DataNode on New Node

- Login to new node
- su hadoop or ssh -X hadoop@192.168.1.103
- Start HDFS on a newly added slave node by using the following command
- ./bin/hadoop-daemon.sh start datanode

Start the DataNode on New Node

- Check the output of jps command on a new node. It looks as follows.
- $ jps
- 7141 DataNode
- 10312 Jps

That’s all for now…