Introduction to Big Data

ECAP456

Dr. Rajni Bhalla

Associate Professor

Learning Outcomes

After this lecture, you will be able to

- understand how to add a new data node in the hadoop cluster
- learn adding user and ssh access.

Starting Hadoop Services

The following command is to start all the Hadoop services on the Hadoop-Master

Adding a New DataNode in The Hadoop Cluster

- Given below are the steps to be followed for adding new nodes to a Hadoop cluster.

Networking

- Add new nodes to an existing Hadoop cluster with some appropriate network configuration.

Adding a New DataNode in The Hadoop Cluster

Assume the following network configuration.

For New node Configuration –

Adding User and SSH Access

Add a User

On a new node, add "hadoop" user and set password of Hadoop user to "hadoop123" or anything you want.

Adding User and SSH Access

- Setup Password less connectivity from master to new slave.
- Execute the following on the master

Adding User and SSH Access

- Execute the following on the 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

Adding User and SSH Access

- Execute the following on the slaves

Login to hadoop. If not, login to hadoop user.

Adding User and SSH Access

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

Adding User and SSH Access

Check ssh login from the master machine. Now check if you can ssh to the new node without a password from the master.

Set Hostname of New Node

You can set hostname in file /etc/sysconfig/network

Set Hostname of New Node

- 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

Set Hostname of New Node

- Update /etc/hosts on all machines of the cluster with the following lines –

Set Hostname of New Node

- Now try to ping the machine with hostnames to check whether it is resolving to IP or not.
- On new node machine –

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.
- Login to new node

Start the DataNode on New Node

- Start HDFS on a newly added slave node by using the following command

Start the DataNode on New Node

- Check the output of jps command on a new node. It looks as follows.


That's all for now...