Hadoop notes

Hadoop installation steps

Install VMWare workstation (or equivalent) and start it

Create new virtual machine:

Select typical installation option

Select option I will install operating system later

Select a name for virtual machine

Selected 20gb as maximum disk capacity

Select option store virtual disk as a single file

Incrased ram allocation to this vm (8gb )

Selected cd/dvd option to point to ubuntu server iso image file which was downloaded into your machine

Click finish

Above created a blank virtual machine.

3. Install Ubuntu server in this virtual machine

1 power on the virtual machine

Import screens men options during ubuntu installation process

1. Do not select keyboard detection
2. Host name select “Hadoop” as the host machine
3. User name select “Hadoop” as the user name
4. Password select “123456” as password
5. DO NOT ENCCRYPT your home directory
6. Partitioning the disk select guided disk option first option
7. Select save changes to the disk Yes
8. HTTP proxy information leave it blank
9. Managing upgrades not automatic
10. Use space bar to select ssh server (first option)
11. Install grub boot loader …Yes

Now that complete log in with username Hadoop and password 123456

Check the ipaddress of your new VM

Use command ifconfig….192.168.11.128

4. Update os

Sudo apt-get update

5. install OpenSSH Server

Sudo

8. Extract the tar file

Tar -xvzf Hadoop-1.2.1.tar.gz

9. Copy the contents of the extracted folder Hadoop 1.2.1 into the location /usr/local/Hadoop

Directory conventions

Windows Linux

Root dir C:\ /

Software c:\Program files /bin

Framework c\windows\System32 /usr/local

Sudo cp -4 hadoop-1.2.1 /usr/local/Hadoop

10. inform system (Linux)wher is Hadoop

Sudo vi /home/Hadoop/.bashrc

{to edit the file in vi editor… press insert}

After fi add:

export HADO\_PREFIX=usr/local/Hadoop

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

Use insert to get into edit mode

escape makes

to save the file and quit escape and :wq

sudo nano /home/Hadoop/.bashrc

(go do bottom and add exports at bottom and use menu options to save and escape)

11. Update the bash settings

Exec bash

1. Inform Hadoop where is java installed on your system

In winscp:

Go to usr.local Hadoop

Look in bin for binary files

Lib folders contains all the jar files

Conf configuration files to configure Hadoop environment

Hadoop-env-sh (click on it)

Set environment variables in it

Set JAVA-HOME variable to where you have set up java.

Open editor in linux to change Hadoop-env.sh file

Sudo vi /user/local/Hadoop/conf/Hadoop-env.sh

(uncomment the java\_home variable and change to):

Export JAVA\_HOME=/user/lib/jvm/java-1.7.0-openjdk-i386

(ensure this path is current)

Export JAVA\_HOME=usr/lib/jvm/java-1.7.0-openjdk-i385

{to save and quit esc :wq }

At this stage you have successfully installed Hadoop in standalone mode

Type Hadoop to get help commands to know if you have installed correctly

Ubuntu 16.04 to use jdk 8

6/14/17

Hadoop Services or (Daemons): provides the capabilities of the Hadoop framework (master/slave architecture)

HDFS

1. NameNode HDFS Master services
2. DataNode HDFS Slave services

MapReduce (Gen1) & Yarn (Yet Another Resource Negotiator) (Gen 2)

1. Jobtracker MapReduce Master services (Gen1 were all MapReduce, but Gen 2 does others)

In Gen 2 in Yard Yet Another Resource Manager Jobtracker is ResourceManager and TraskTracker is called NodeManager

1. Tasktracker MapReduce Slave services

Checkpoint Services

1. Secondary Namenode

MapReduce is not replaced by YARN, MapReduce is the application framework and distributed processing architecture. With YARN the application framework is decoupled from the processing architecture by use of YARN, it is a newer version of MapReduce. The only difference is that different processing can be done on the same architecture

Example: 10 node system: 1 would be name node and others would provide data node services

Hadoop Executor Modes

Hadoop Service acronyms: Name Node – NM, Data Node - DN Job tracker – JT, Task Tracker TT Secondary Name Node SNM

The way your are starting the cluster are changed by configuration and number of computers on cluster

1 Standalone mode: one physical machine with a virtual machine that is a single node cluster with name node, data node, job tracker, tasktracker and secondary NM all running on one machine. No HDFS available.

Basic installation of Hadoop

* All 5 Hadoop services run on a single node (machine and within single JVM (single process)
* HDFS (Hadoop Distributed Files System) is NOT present (MapReduce etc that you run on the closer will be reading and writing on to local file system.)
* Hadoop uses local file system (Linux)

2 Psuedo-distributed mode: Single machine but all of 5 services are running as independent virtual processes on the same machine. And you do have HDFS. All services – separate JVMs

* All 5 Hadoop services run on a single node but as separate processes (JVMs)
* HDFS is present
* Not fully distributed because it is still a single machine

3 fully-distributed mode: more than one machine in the services. Name Node server with Name Node and Job tracker is on a dedicated machine. Secondary name node is on a dedicated machine called secondary name node server. All other nodes have data node and task tracker running on them. HDFS is present

* Master(NameNode, Jobtracker/Resource Manager) and checkpoint service (Secondary Name node) Slave services (DataNode, tasktracker or node manager) run on separate nodes together forming Hadoop cluster

Call have 100 node will have 98 data nodes on it. Two nodes reserved for master and check point services

Limitation: in Gen 1 n = 4096 – 2

Gen2 n = unlimited

Edge nodes are an interface between Hadoop cluster and outside nodes

Setup Hadoop in Pseudo-distributed mode

In order to set Hadoop in pseudo-distributed mode, you need ot perform all stesp (1 to 12) that we used to set up Hadoop in Standalone mode and then

1. Configure all the 5 Hadoop serivedes and HDFS
2. 2. Create HDFS by formatting the NameNode
3. 3. Start all the Hadoop services
4. Setup the name node ser vices

* Open the file core-site. Xml to configure NameNode service

Set 2 properties

Fs.default.name ----🡪 system: portnumber where namenode service will be started.

Hadoop.tmp.dir ---🡪 name of folder to be used by HDFS for internal storage

<property>

<name>fs.default.name</name>

<value>hdfs://192.168.11.129:10001</value>

</property>

<property>

<name>Hadoop.tmp.dir</name>

<value>/usr/

14/ Set up the job tracker service

* Open the file mapred-site.xml to configure jobtracker

GET HIS NOTES!!!!

NM 50070

JT 50030

TT 500

Putty edition : ??? Study putty and using putty