

Journal

Date: 25/11/2020

Did small research to decide which technology to use and what other supporting technology to add that goes along with it.

Date: 02/12/2020

Decided to do Apache Hadoop and Apache Pig both are used in processing huge data and the Pig uses SQL like language Pig Latin and uses the MapReduce to give quick result, to deploy the Vm's Vagrant and for VM's

Date: 11/12/2020

Finding a suitable data to process and use hive to perform HDFS and MapReduce

Date: 15/12/2020

Implementing

Used Vagrant to code and deploy 3 VM's ping each other to re-check the validity of IP and to check if ssh were working.

Node04, Node02, Node03

Date: 16/12/2021

Installed java, begin to install Apache Hadoops on all 3 nodes. Decided to make Node 01 as master and the rest of the nodes as slaves.

For the password less login b/w name and data nodes ssh keygen is used and the id_rsa.pub keys are pasted into the authorized keys for easy accesses between the master and slave.

Downloaded the Hadoop-3.1.1 from the Apache website and unzipped the file using 'tar -xzf' in to separate Hadoop user.

Updated the core-site.xml and hdfs-site.xml as per the name and data nodes

Date:19/12/2021

The workser file in Hadoop was updated with IP of slave nodes and before starting the Hadoop cluster the hdfs was formatted and the user Hadoop could not create the file data cause the permission was denied, manually adder the data file with mkdir and chmod 777 was used to solve this issue and any further issue.

The HDFS service was started but the nodes couldn't communicate with each other cause of a public key error and permission was denied.

Date:20/12/2021

The reason for it was the mismatch in the ssh keys redid the ssh keygen and copied the keys to the authorized keys. But this didn solved the issue.

Date:21/12/2021

The Hadoop user was given sudo privilege by adding it to sudoers to check if that would solve the issue, but the issue was solved by editing the "sshd_config" like changing the PubKeyAuthentication Yes and setting UsePAM yes.

```
hadoop@node04: ~  
has been successfully formatted.  
2021-12-23 14:49:07,258 INFO namenode.FSImageFormatProtobuf: Saving image file /usr/local/hadoop  
p/data/nameNode/current/fsimage.ckpt_000000000000000000 using no compression  
2021-12-23 14:49:07,469 INFO namenode.FSImageFormatProtobuf: Image file /usr/local/hadoop/data/  
nameNode/current/fsimage.ckpt_000000000000000000 of size 401 bytes saved in 0 seconds .  
2021-12-23 14:49:07,487 INFO namenode.NNStorageRetentionManager: Going to retain 1 images with  
txid >= 0  
2021-12-23 14:49:07,515 INFO namenode.FSNamesystem: Stopping services started for active state  
2021-12-23 14:49:07,522 INFO namenode.FSNamesystem: Stopping services started for standby state  
2021-12-23 14:49:07,533 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet  
shutdown.  
2021-12-23 14:49:07,533 INFO namenode.NameNode: SHUTDOWN_MSG:  
/*****  
SHUTDOWN_MSG: Shutting down NameNode at node04/198.16.0.40  
*****/  
hadoop@node04:~$ start-dfs.sh  
Starting namenodes on [node04]  
Starting datanodes  
Starting secondary namenodes [node04]  
hadoop@node04:~$ jps  
7938 NameNode  
8284 Jps  
8190 SecondaryNameNode  
hadoop@node04:~$
```

Date:22/12/2021

Cluster started, but when accessed through the web UI it showed 0 nodes, and there were no Node in master and no data node in the save nodes.

This error was due to different permission the existed between the name and data node files.

Chmod755, and 700 were given respectively for the data and name node this issue was solved and the Web UI for the same was also working.

The screenshot shows the Hadoop DFS Health web UI in a browser. The address bar shows the URL: 198.16.0.40:9870/dfshealth.html#tab-overview. The page title is "Overview 'node04:9000' (active)".

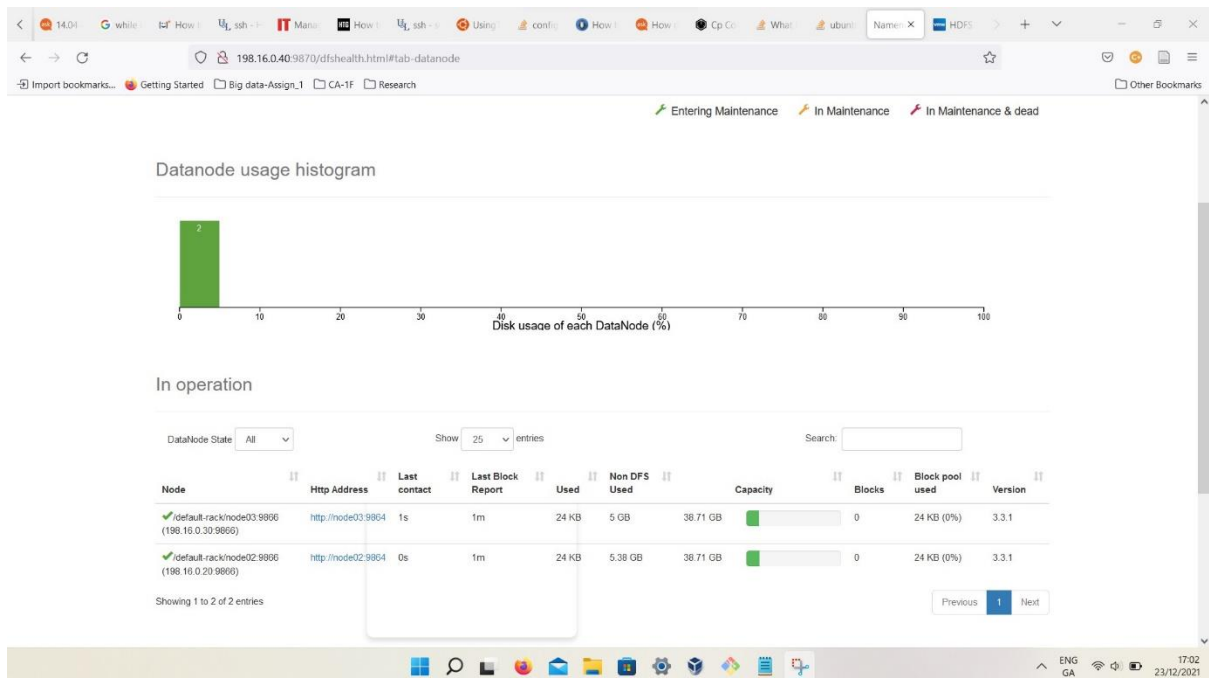
Overview 'node04:9000' (active)

Started:	Thu Dec 23 16:58:31 +0000 2021
Version:	3.3.1, ra309c37a397ad4168041d80521bdeefc45685f2
Compiled:	Tue Jun 15 06:13:00 +0100 2021 by ubuntu from (HEAD detached at release-3.3.1-RC3)
Cluster ID:	CID-8330190-5001-422a-b495-4c0c209db779
Block Pool ID:	BP-221430756-198.16.0.40-1640276951730

Summary

Security is off.
Safemode is off.
1 files and directories, 0 blocks (0 replicated blocks, 0 erasure coded block groups) = 1 total filesystem object(s).
Heap Memory used 25.01 MB of 42.14 MB Heap Memory. Max Heap Memory is 243.63 MB.
Non-Heap Memory used 50.78 MB of 52.13 MB Committed Non-Heap Memory. Max Non-Heap Memory is <unbounded>.

Configured Capacity:	77.43 GB
Configured Remote Capacity:	0 B
DFS Used:	48 KB (0%)
Non DFS Used:	10.38 GB
DFS Remaining:	67.01 GB (86.95%)
Block Pool Used:	48 KB (0%)
DataNodes usages% (Min/Median/Max/stdDev):	0.00% / 0.00% / 0.00% / 0.00%
Live Nodes	2 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)



Mapre-site.xml and yarn-site.xml were edited to start the yarn scheduler service was started and was verified by checking the yarn node -list.

How to install hadoop cluster/multi node on ubuntu server 18.04

8,454 views • Oct 7, 2019

The video content shows a terminal window with the following commands and output:

```

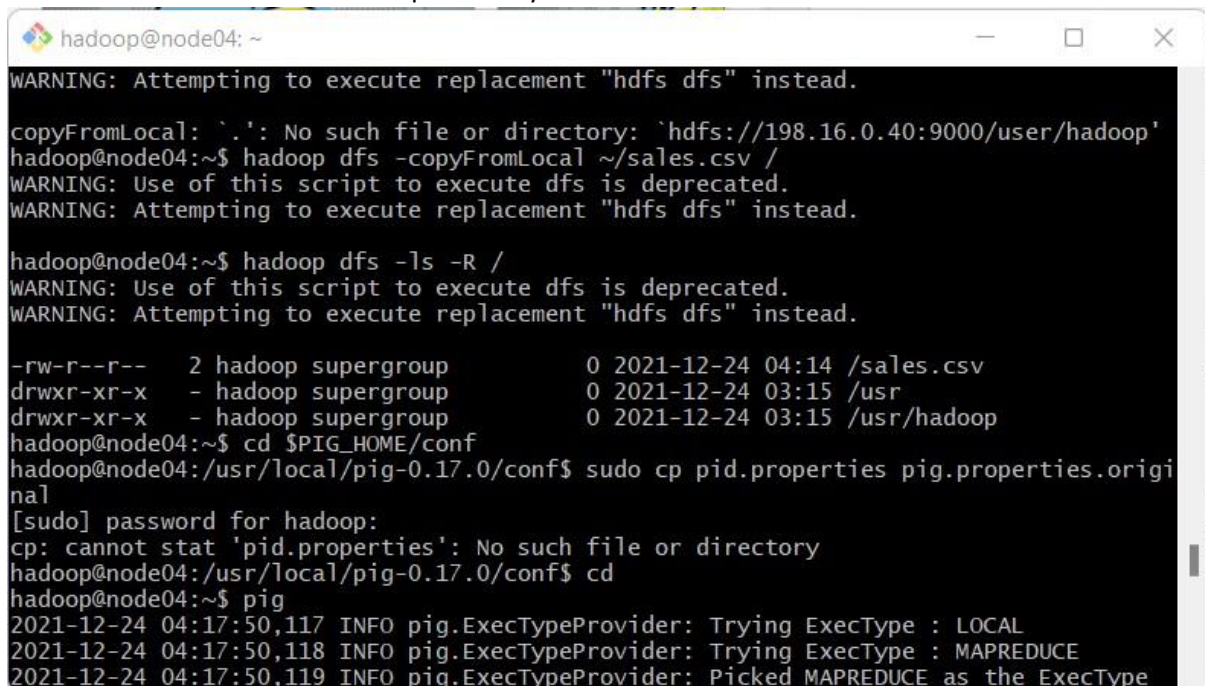
$ start-yarn.sh
$ start-resource-manager
$ start-namenode
$ yarn node -list
2021-12-22 17:12:49,784 INFO client.DefaultHARMAFailoverProxyProvider: Connecting to Resource
anager at /198.16.0.40:8032
Total Nodes:2
Node-Id Node-Http-Address Node-State Node-Http-Address Number-of-Running-Containers
node03:42783 RUNNING node03:8042 0
node02:39217 RUNNING node02:8042 0
hadoopnode04:~$

```

There are safety running containers because of the...

Date:23/12/2021

Installation of Apache Pig, the file was downloaded from the Apache pig and was extracted with tar -xvzf and was moved to the Hadoop directory.

A terminal window titled 'hadoop@node04: ~' showing a series of commands and their outputs. The commands include 'hadoop dfs -copyFromLocal ~/sales.csv /', 'hadoop dfs -ls -R /', and 'pig'. The outputs show warnings about deprecated scripts and file permissions, followed by a directory listing of /usr/hadoop, and then Pig execution logs indicating the selection of MAPREDUCE as the execution type.

```
hadoop@node04: ~  
WARNING: Attempting to execute replacement "hdfs dfs" instead.  
copyFromLocal: `.`: No such file or directory: `hdfs://198.16.0.40:9000/user/hadoop'  
hadoop@node04:~$ hadoop dfs -copyFromLocal ~/sales.csv /  
WARNING: Use of this script to execute dfs is deprecated.  
WARNING: Attempting to execute replacement "hdfs dfs" instead.  
  
hadoop@node04:~$ hadoop dfs -ls -R /  
WARNING: Use of this script to execute dfs is deprecated.  
WARNING: Attempting to execute replacement "hdfs dfs" instead.  
  
-rw-r--r--    2 hadoop supergroup          0 2021-12-24 04:14 /sales.csv  
drwxr-xr-x    - hadoop supergroup          0 2021-12-24 03:15 /usr  
drwxr-xr-x    - hadoop supergroup          0 2021-12-24 03:15 /usr/hadoop  
hadoop@node04:~$ cd $HADOOP_HOME/conf  
hadoop@node04:/usr/local/pig-0.17.0/conf$ sudo cp pid.properties pig.properties.origi  
nal  
[sudo] password for hadoop:  
cp: cannot stat 'pid.properties': No such file or directory  
hadoop@node04:/usr/local/pig-0.17.0/conf$ cd  
hadoop@node04:~$ pig  
2021-12-24 04:17:50,117 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL  
2021-12-24 04:17:50,118 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE  
2021-12-24 04:17:50,119 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
```

Using Pig Latin, a data set for the sales were analyzed, the code used are as below:

First the data was copied from local to HDFS, was moved into the directory

```
$HADOOP_HOME/bin/hdfs dfs -copyFromLocal file:///C:/Download/Sale.csv /
```

And the csv file

By executing 'pig' apache Pig starts the service.

```
grunt> salesTable = LOAD '/SalesJan2009.csv' USING PigStorage(',') AS  
(Transaction_date:chararray,Product:chararray,Price:chararray,Payment_Type:  
chararray,Name:chararray,City:chararray,State:chararray,Country:chararray,A  
ccount_Created:chararray,Last_Login:chararray,Latitude:chararray,Longitude:  
chararray);
```

```
GroupByCountry = GROUP salesTable BY Transaction_Date;
```

```
Transaction_DateByCountry = FOREACH GroupByCountry GENERATE  
CONCAT((chararray)$0,CONCAT(':', (chararray)COUNT($1)));
```

```
STORE Transaction_DateByCountry INTO 'pig_output_sales' USING  
PigStorage('\t');
```

```
store pig_output_sales into '/user/hadoop/csvoutput' using  
PigStorage('\t','-schema');
```

```
hadoop fs -getmerge /user/hadoop/csvoutput ./output.csv
```

```
hadoop@node04: ~
2021-12-24 04:17:50,669 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://198.16.0.40:9000
2021-12-24 04:17:51,260 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: 198.16.0.40:54311
2021-12-24 04:17:51,293 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-760771a7-0cef-45df-a048-ef88c669e1d3
2021-12-24 04:17:51,293 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> salesTable = LOAD '/sales.csv' USING PigStorage(',') AS (Transaction_date:chararray,Product:chararray,Price:chararray,Payment_Type:chararray,Name:chararray,City:chararray,State:chararray,Country:chararray,Account_Created:chararray,Last_Login:chararray,Latitude:chararray,Longitude:chararray);
grunt> GroupByCountry = GROUP salesTable BY Country;
grunt> CountByCountry = FOREACH GroupByCountry GENERATE CONCAT((chararray)$0,CONCAT(':', (chararray)COUNT($1)));
grunt> STORE CountByCountry INTO 'pig_output_sales' USING PigStorage('\t');
2021-12-24 04:20:25,162 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.output.textoutputformat.separator
2021-12-24 04:20:25,209 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: GROUP_BY
2021-12-24 04:20:25,242 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2021-12-24 04:20:25,274 [main] INFO org.apache.pig.newplan.logical.optimizer.Logical
```