Install hadoop in linux (Ubuntu)

Install JDK on Ubuntu:

$ sudo apt update

$ sudo apt install openjdk-8-jdk -y

- Once the installation píocess is complete, veíify the cuíient Java veísion:

$ java -version; javac -version

```
pnap@phoenixnap:~$ java -version; javac -version
openjdk version "1.8.0_422"
OpenJDK Runtime Environment (build 1.8.0_422-8u422-b05-1~24.04-b05)
OpenJDK 64-Bit Server VM (build 25.422-b05, mixed mode)
javac 1.8.0_422
```

Install OpenSSH on Ubuntu

$ sudo apt install openssh-server openssh-client -y

- In the example below, the output confirms that the latest version isalready installed.

```
pnap@phoenixnap:~$ sudo apt install openssh-server openssh-client -y
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
openssh-server is already the newest version (1:9.6p1-3ubuntu13.5).
openssh-client is already the newest version (1:9.6p1-3ubuntu13.5).
openssh-client set to manually installed.
0 upgraded, 0 newly installed, 0 to remove and 166 not upgraded.
```

Create Hadoop User

$ sudo adduser hdoop

- The username, in this example, is **hdoop**. You are free to use any username and password you see fit.

- Switch to the newly created user and enter the corresponding password:

$su – hdoop

Enable Passwordless SSH for Hadoop User

$ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa

```
hdoop@phoenixnap:~$ ssh-keygen -t rsa -P '' -f ~/.ssh/id_rsa
Generating public/private rsa key pair.
Created directory '/home/hdoop/.ssh'.
Your identification has been saved in /home/hdoop/.ssh/id_rsa
Your public key has been saved in /home/hdoop/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:DFtcZg3wmo56IKQKdGTWSG8/+YePol1UvGWpVPpoy34 hdoop@phoenixnap
The key's randomart image is:
+---[RSA 3072]----+
|  ..o   ..=o .   |
|   =.. . =. + .  |
|  +  o. o .= +   |
| .... .=.oo B    |
|.o.   .+S. = .   |
|o . .  o+ + .    |
|o  . .. .+ +     |
|.    .o.. =  E   |
|    .o.... o.    |
+----[SHA256]-----+
hdoop@phoenixnap:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
hdoop@phoenixnap:~$ chmod 0600 ~/.ssh/authorized_keys
```

$cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys

$ chmod 0600 ~/.ssh/authorized_keys

$ssh localhost

Download and Install Hadoop on Ubuntu

$ wget https://dlcdn.apache.org/hadoop/common/hadoop-3.4.0/hadoop-3.4.0.tar.gz

```
hdoop@phoenixnap:~$ wget https://dlcdn.apache.org/hadoop/common/hadoop-3.4.0/hadoop
-3.4.0.tar.gz
--2024-09-09 11:53:23--  https://dlcdn.apache.org/hadoop/common/hadoop-3.4.0/hadoop
-3.4.0.tar.gz
Resolving dlcdn.apache.org (dlcdn.apache.org)... 151.101.2.132, 2a04:4e42::644
Connecting to dlcdn.apache.org (dlcdn.apache.org)|151.101.2.132|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 965537117 (921M) [application/x-gzip]
Saving to: ‘hadoop-3.4.0.tar.gz’

hadoop-3.4.0.tar.gz            100%[=================================================
=====>] 920.81M  3.29MB/s    in 4m 31s

2024-09-09 11:57:55 (3.39 MB/s) - ‘hadoop-3.4.0.tar.gz’ saved [965537117/965537117]
```

$ tar xzf hadoop-3.4.0.tar.gz

$nano .bashrc

-insert this : #Hadoop Related Options

export HADOOP_HOME=/home/hdoop/hadoop-3.4.0 export HADOOP_INSTALL=$HADOOP_HOME

export HADOOP_MAPRED_HOME=$HADOOP_HOME export HADOOP_COMMON_HOME=$HADOOP_HOME export HADOOP_HDFS_HOME=$HADOOP_HOME

export YARN_HOME=$HADOOP_HOME

export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin

export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"

```
  GNU nano 7.2                                    .bashrc *
  if [ -f /usr/share/bash-completion/bash_completion ]; then
    . /usr/share/bash-completion/bash_completion
  elif [ -f /etc/bash_completion ]; then
    . /etc/bash_completion
  fi
fi

#Hadoop Related Options
export HADOOP_HOME=/home/hdoop/hadoop-3.4.0
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"

^G Help        ^O Write Out   ^W Where Is    ^K Cut         ^T Execute
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify
```

-click Ctrl+s and Ctrl+x to go out the editor

$source ~/.bashrc

#Edit hadoop-env.sh File

$nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh

- Uncomment the **$JAVA_HOME** variable (i.e., remove the # sign) and add the full path to the OpenJDK installation on your system. If you haveinstalled the same version as presented in the first part of this tutorial,add the following line:

$export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

```
  GNU nano 7.2              /home/hdoop/hadoop-3.4.0/etc/hadoop/hadoop-env.sh *

# Technically, the only required environment variable is JAVA_HOME.
# All others are optional.  However, the defaults are probably not
# preferred.  Many sites configure these options outside of Hadoop,
# such as in /etc/profile.d

# The java implementation to use. By default, this environment
# variable is REQUIRED on ALL platforms except OS X!
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

# The language environment in which Hadoop runs. Use the English
# environment to ensure that logs are printed as expected.
export LANG=en_US.UTF-8

# Location of Hadoop.  By default, Hadoop will attempt to determine
# this location based upon its execution path.
# export HADOOP_HOME=


^G Help        ^O Write Out   ^W Where Is    ^K Cut         ^T Execute
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify
```

-click Ctrl+s and Ctrl+x to go out the editor

#Edit core-site.xml File

Note: keep this number , we will use it alot

$hostname -I | awk '{print $1}'

-copy the number address

$nano $HADOOP_HOME/etc/hadoop/core-site.xml

-Add the following configuration to override the default values for the

temporary directory and add your HDFS URL to replace the default localfile system setting:

<configuration>

<property>

```
<name>hadoop.tmp.dir</name>

<value>/home/hdoop/tmpdata</value>
```

</property>

<property>

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

<value>hdfs://put the number here:9000</value>

</property>

</configuration>

```
  GNU nano 7.2              /home/hdoop/hadoop-3.4.0/etc/hadoop/core-site.xml *
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
  <name>hadoop.tmp.dir</name>
  <value>/home/hdoop/tmpdata</value>
</property>
<property>
  <name>fs.default.name</name>
  <value>hdfs://127.0.0.1:9000</value>
</property>
</configuration>

^G Help        ^O Write Out   ^W Where Is    ^K Cut         ^T Execute
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify
```

#**Edit hdfs-site.xml File**

$ sudo nano $HADOOP_HOME/etc/hadoop/hdfs-site.xml

-Add the following configuration to the file and, if needed, adjust the NameNode and DataNode directories to your custom locations:

<configuration>

<property>

```
 <name>dfs.data.dir</name>
 <value>/home/hdoop/dfsdata/namenode</value>
</property>
<property>
 <name>dfs.data.dir</name>
 <value>/home/hdoop/dfsdata/datanode</value>
</property>
<property>
 <name>dfs.replication</name>
 <value>1</value>
</property>
</configuration>
```

Note: -make sure you replace “hdoop” with your user, you can find it in the command line,like here :

```
hdoop@NuvobookV1:~$
```

```
  GNU nano 7.2              /home/hdoop/hadoop-3.4.0/etc/hadoop/hdfs-site.xml *

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
  <name>dfs.data.dir</name>
  <value>/home/hdoop/dfsdata/namenode</value>
</property>
<property>
  <name>dfs.data.dir</name>
  <value>/home/hdoop/dfsdata/datanode</value>
</property>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
</configuration>

^G Help        ^O Write Out   ^W Where Is    ^K Cut         ^T Execute
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify
```

#**Edit mapred-site.xml File**

$sudo nano $HADOOP_HOME/etc/hadoop/mapred-site.xml

-Add the following configuration to change the default MapReduceframework name value to **yarn**:

```
<configuration>

<property>

 <name>mapreduce.framework.name</name>

 <value>yarn</value>

</property>
```

```
</configuration>
```

```
  GNU nano 7.2            /home/hdoop/hadoop-3.4.0/etc/hadoop/mapred-site.xml *
    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
  <name>mapreduce.framework.name</name>
  <value>yarn</value>
</property>
</configuration>

^G Help        ^O Write Out   ^W Where Is    ^K Cut         ^T Execute
^X Exit        ^R Read File   ^\ Replace     ^U Paste       ^J Justify
```

Edit yarn-site.xml File

$nano $HADOOP_HOME/etc/hadoop/yarn-site.xml

-Append the following configuration to the file:

```
<configuration>

<property>

 <name>yarn.nodemanager.aux-services</name>

 <value>mapreduce_shuffle</value>

</property>

<property>

 <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
```

```
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
```

```
</property>

<property>

 <name>yarn.resourcemanager.hostname</name>

 <value>put the same number that you puut before here</value>

</property>

<property>

 <name>yarn.acl.enable</name>

 <value>0</value>

</property>

<property>

 <name>yarn.nodemanager.env-whitelist</name>


<value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HA
DOOP_CONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HO
ME,HADOOP_MAPRED_HOME</value>

</property>

</configuration>
```

```
GNU nano 7.2                    /home/hdoop/hadoop-3.4.0/etc/hadoop/yarn-site.xml *
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->
<configuration>
<property>
  <name>yarn.nodemanager.aux-services</name>
  <value>mapreduce_shuffle</value>
</property>
<property>
  <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
  <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property>
  <name>yarn.resourcemanager.hostname</name>
  <value>127.0.0.1</value>
</property>
<property>
  <name>yarn.acl.enable</name>
  <value>0</value>
</property>
<property>
  <name>yarn.nodemanager.env-whitelist</name>
  <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
</property>
</configuration>

^G Help      ^O Write Out  ^W Where Is   ^K Cut       ^T Execute    ^C Location   M-U Undo      M-A Set Mark  M-] To Bracket
^X Exit      ^R Read File  ^\ Replace    ^U Paste     ^J Justify    ^/ Go To Line M-E Redo      M-6 Copy      ^Q Where Was
```

Format HDFS NameNode

$hdfs namenode -format

-The shutdown notification signifies the end of the NameNode formatprocess.

```
hdoop@phoenixnap:~$ hdfs namenode -format
WARNING: /home/hdoop/hadoop-3.4.0/logs does not exist. Creating.
2024-09-09 13:08:42,739 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = phoenixnap/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.4.0
STARTUP_MSG:   classpath = /home/hdoop/hadoop-3.4.0/etc/hadoop:/home/hdoop/hado
3.4.0/share/hadoop/common/lib/curator-client-5.2.0.jar:/home/hdoop/hadoop-3.4.0,
2024-09-09 13:08:45,012 INFO namenode.FSNamesystem: Stopping services started f
standby state
2024-09-09 13:08:45,018 INFO namenode.FSImage: FSImageSaver clean checkpoint: t
=0 when meet shutdown.
2024-09-09 13:08:45,019 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at phoenixnap/127.0.1.1
************************************************************/
```

Start Hadoop Cluster

$cd

$cd *hadoop-3.4.0/sbin*

- execute the following command to start the NameNode andDataNode:

$./start-dfs.sh

```
hdoop@phoenixnap:~/hadoop-3.4.0/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [phoenixnap]
```

- Once the *namenode*, *datanodes*, and *secondary namenode* are upand running, start the YARN resource and *nodemanagers* by typing:

$./start-yarn.sh

```
hdoop@phoenixnap:~/hadoop-3.4.0/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
```

-Run the following command to check if all the daemons are active and running as Java processes:

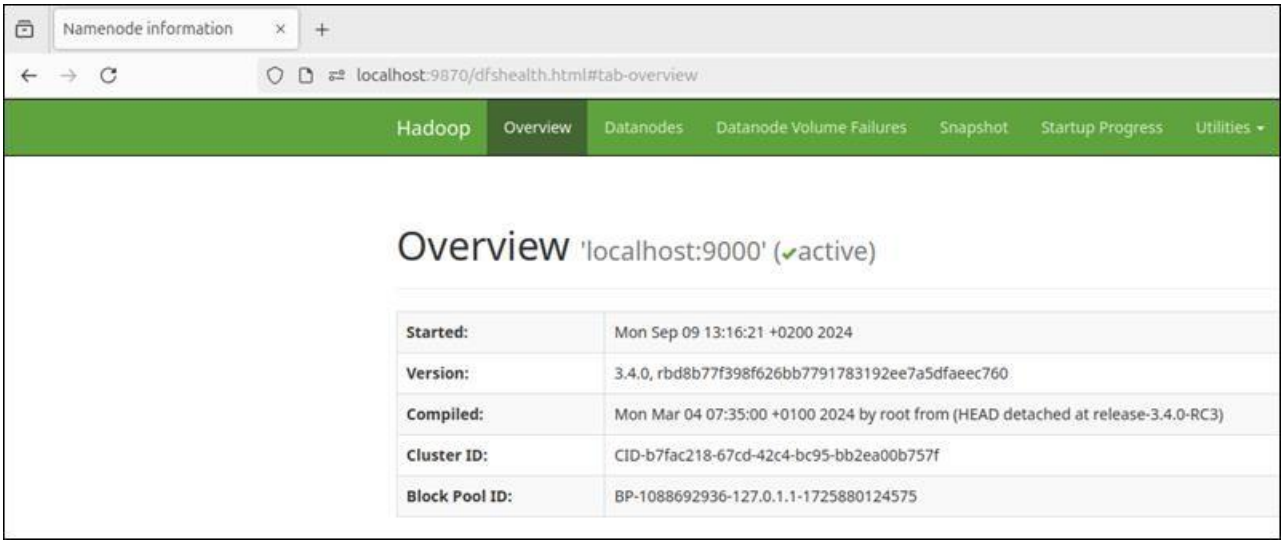
$jps

-If everything works as intended, the resulting list of running Javaprocesses contains all the HDFS and YARN daemons.

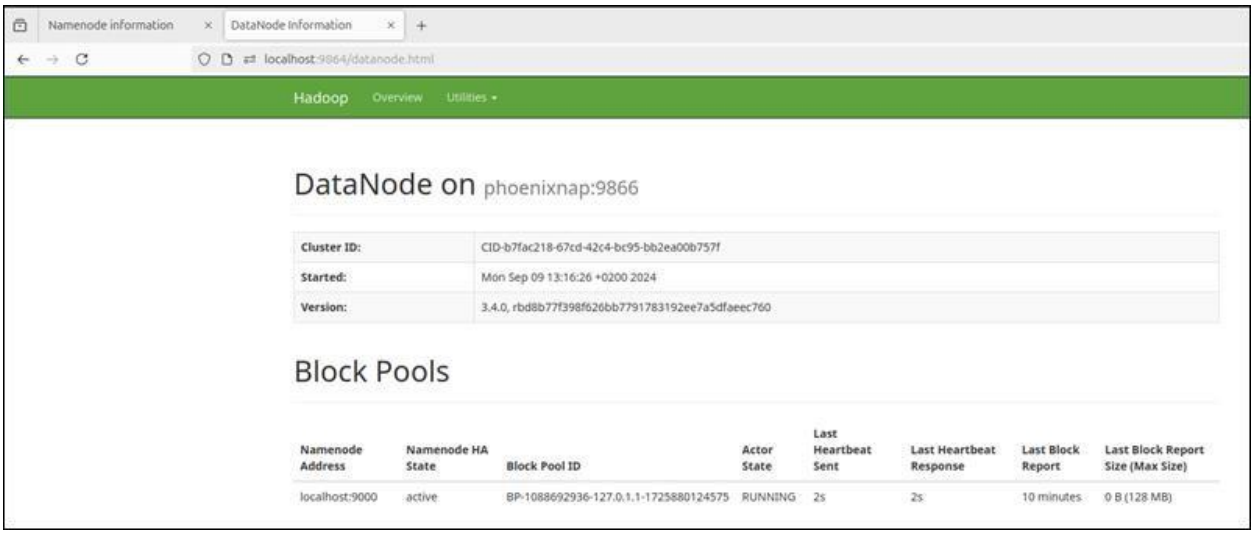
```
hdoop@phoenixnap:~/hadoop-3.4.0/sbin$ jps
45169 DataNode
46355 ResourceManager
45033 NameNode
46476 NodeManager
45373 SecondaryNameNode
47390 Jps
```

Access Hadoop from Browser

#nameNode http://localhost:9870



#dataNode: http://localhost:9864



#resorce manager:

http://_____put_____the same number here:8088

Hive installation in Linux (ubuntu)

#login to hdoop user

$ su – hdoop

```
mdtaha@NuvobookV1:~$ su - hdoop
Password:
hdoop@NuvobookV1:~$
```

#start Hadoop server

$ start-all.sh

$ jps

```
hdoop@NuvobookV1:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hdoop in 10 second
s.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [172.18.132.51]
Starting datanodes
Starting secondary namenodes [NuvobookV1]
Starting resourcemanager
Starting nodemanagers
hdoop@NuvobookV1:~$ jps
13749 NodeManager
13627 ResourceManager
13163 DataNode
13405 SecondaryNameNode
14158 Jps
13039 NameNode
hdoop@NuvobookV1:~$
```

#download hive 3.1.3 and extract it

$ wget http://download.nust.na/pub2/apache/hive/hive-3.1.3/apache-hive- 3.1.3-bin.tar.gz

```
hdoop@phoenixNAP:~$ wget https://downloads.apache.org/hive/hive-4.0.0/apache-hiv
e-4.0.0-bin.tar.gz
--2024-09-02 07:57:53--  https://downloads.apache.org/hive/hive-4.0.0/apache-hiv
e-4.0.0-bin.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.208.237, 135.181.
214.104, 2a01:4f8:10a:39da::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|88.99.208.237|:443...
connected.
HTTP request sent, awaiting response... 200 OK
Length: 458782861 (438M) [application/x-gzip]
Saving to: 'apache-hive-4.0.0-bin.tar.gz'

apache-hive-4.0.0-b 100%[===================>] 437.53M  17.6MB/s    in 26s

2024-09-02 07:58:19 (16.9 MB/s) - 'apache-hive-4.0.0-bin.tar.gz' saved [45878286
1/458782861]
```

```
hdoop@NuvobookV1:~$ tar xzf apache-hive-3.1.3-bin.tar.gz
hdoop@NuvobookV1:~$ ls
apache-hive-3.1.3-bin          hadoop-3.4.0                pig-0.17.0
apache-hive-3.1.3-bin.tar.gz   hadoop-3.4.0.tar.gz         pig-0.17.0.tar.gz
derby.log                      hadoop-streaming-2.7.3.jar  sample.txt
dev                            lab                         test.txt
dfsdata                        metastore_db                tmpdata
```

$ tar xzf apache-hive-3.1.3-bin.tar.gz

#configur system variables

$nano .bashrc

-insert this:

export HIVE_HOME=/home/hdoop/apache-hive-3.1.3-bin export PATH=$PATH:$HIVE_HOME/bin

export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64 export PATH=$PATH:$JAVA_HOME/bin

```
#Hive sitting
export HIVE_HOME=/home/hdoop/apache-hive-3.1.3-bin
export PATH=$PATH:$HIVE_HOME/bin
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export PATH=$PATH:$JAVA_HOME/bin
```

-click Ctrl+s and Ctrl+x to go out the editor

$ source ~/.bashrc

```
hdoop@NuvobookV1:~$ nano .bashrc
hdoop@NuvobookV1:~$ source .bashrc
hdoop@NuvobookV1:~$
```

#Make directories inside Hadoop and change their permissions:

$hadoop fs -mkdir /tmp

$hadoop fs -chmod g+w /tmp

```
hdoop@NuvobookV1:~$ hadoop fs -mkdir /tmp
hdoop@NuvobookV1:~$ hadoop fs -chmod g+w /tmp
hdoop@NuvobookV1:~$
```

$hadoop fs -mkdir /user

$hadoop fs -mkdir /user/hive

$hadoop fs -mkdir /user/hive/warehouse

$hadoop fs -chmod g+w /user/hive/warehouse

```
hdoop@NuvobookV1:~$ hadoop fs -mkdir /user
hdoop@NuvobookV1:~$ hadoop fs -mkdir /user/hive
hdoop@NuvobookV1:~$ hadoop fs -mkdir /user/hive/warehouse
hdoop@NuvobookV1:~$ hadoop fs -chmod g+w /user/hive/warehous
```

Edit hive-config.sh file

export HADOOP_HOME=/home/hdoop/hadoop-3.4.0

```
# Disable the JNDI. This feature has critical RCE vulnerability.
# when 2.x <= log4j.version <= 2.14.1
export HADOOP_CLIENT_OPTS="$HADOOP_CLIENT_OPTS -Dlog4j2.formatMsgNoLookups=>
export HADOOP_HOME=/home/hdoop/hadoop-3.4.0
```

-clickCtrl+s and Ctrl+x to go out the editor

#replace guava file

$ cd sudo $HIVE_HOME/lib/

$ ls -l guava*

$ rm -rf guava-19.0.jar

$ cp $HADOOP_HOME/share/hadoop/hdfs/lib/guava-27.0-jre.jar

$HIVE_HOME/lib

```
hdoop@NuvobookV1:~$ cd $HIVE_HOME/lib/
hdoop@NuvobookV1:~/apache-hive-3.1.3-bin/lib$ ls -l guava*
-rw-r--r-- 1 hdoop hdoop 2308517 Oct 18  2019 guava-19.0.jar
hdoop@NuvobookV1:~/apache-hive-3.1.3-bin/lib$ rm -rf guava-19.0.jar
hdoop@NuvobookV1:~/apache-hive-3.1.3-bin/lib$ ls -l guava*
ls: cannot access 'guava*': No such file or directory
hdoop@NuvobookV1:~/apache-hive-3.1.3-bin/lib$  cp $HADOOP_HOME/share/hadoop/
hdfs/lib/guava-27.0-jre.jar $HIVE_HOME/lib
hdoop@NuvobookV1:~/apache-hive-3.1.3-bin/lib$ ls -l guava*
-rw-r--r-- 1 hdoop hdoop 2747878 Oct 17 08:00 guava-27.0-jre.jar
hdoop@NuvobookV1:~/apache-hive-3.1.3-bin/lib$
```

Configure hive-site.xml File

$ cd $HIVE_HOME/conf

$ ls -l

```
hdoop@phoenixNAP:~/apache-hive-4.0.0-bin/conf$ ls -l
total 844
-rw-r--r-- 1 hdoop hdoop   1775 Jan 22  2020 beeline-log4j2.properties.template
-rw-r--r-- 1 hdoop hdoop 413104 Jan 22  2020 hive-default.xml.template
-rw-r--r-- 1 hdoop hdoop   2365 Jan 22  2020 hive-env.sh.template
-rw-r--r-- 1 hdoop hdoop   2274 Jan 22  2020 hive-exec-log4j2.properties.template
-rw-r--r-- 1 hdoop hdoop   3086 Jan 22  2020 hive-log4j2.properties.template
-rw-r--r-- 1 hdoop hdoop 413104 Sep  3 04:15 hive-site.xml
-rw-r--r-- 1 hdoop hdoop   2060 Jan 22  2020 ivysettings.xml
-rw-r--r-- 1 hdoop hdoop   3558 Jan 22  2020 llap-cli-log4j2.properties.template
-rw-r--r-- 1 hdoop hdoop   7093 Jan 22  2020 llap-daemon-log4j2.properties.template
-rw-r--r-- 1 hdoop hdoop   2662 Jan 22  2020 parquet-logging.properties
```

$ cp hive-default.xml.template hive-site.xml

```
$ nano hive-site.xml
```

-click Ctrl+w to search for "**hive.metastore.warehouse.dir** ", do it againand again until you find this

```
GNU nano 7.2                    hive-site.xml
      Type of database used by the metastore. Information schema &amp; JDBCStor>
    </description>
  </property>
  <property>
    <name>hive.metastore.warehouse.dir</name>
    <value>/user/hive/warehouse</value>
    <description>location of default database for the warehouse</description>
  </property>
```

-add **"/home/hdoop**" , replace hdoop wit your system username.

```
hdoop@NuvobookV1:~$
```

```
    <description>
<ire Exclusive locks for&#8;transactional tables.  This ensures that insert>
      are not hidden by the INSERT OVERWRITE.
    </description>
  </property>
  <property>
    <name>hive.txn.timeout</name>
    <value>300s</value>
    <description>
      Expects a time value with unit (d/day, h/hour, m/min, s/sec, ms/msec,>
      time after which transactions are declared aborted if the client has >
    </description>
  </property>
  <property>
    <name>hive.txn.heartbeat.threadpool.size</name>
    <value>5</value>
    <description>The number of threads to use for heartbeating. For Hive CL>
  </property>
  <property>
    <name>hive.txn.manager.dump.lock.state.on.acquire.timeout</name>
    <value>false</value>
Search [for&#8]:  for&#8
^G Help        M-C Case Sens   M-B Backwards   ^P Older        ^T Go To Line
^C Cancel      M-R Reg.exp.    ^R Replace      ^N Newer
```

-click Ctrl+w and search for **"for&#8"** and remove &#8

-click Ctrl+s and Ctrl+x to go out the editor

```
$ nano hive-site.xml
```

-add these under **“<configuration>”**

<property>

<name>system:java.io.tmpdir</name>

<value>/tmp/hive/java</value>

</property>

<property>

<name>system:user.name</name>

<value>${user.name}</value>

</property>

```
-->
<configuration>
  <!-- WARNING!!! This file is auto generated for documentation purposes ON>
  <!-- WARNING!!! Any changes you make to this file will be ignored by Hive>
  <!-- WARNING!!! You must make your changes in hive-site.xml instead.     >
  <!-- Hive Execution Parameters -->


<property>
 <name>system:java.io.tmpdir</name>
 <value>/tmp/hive/java</value>
 </property>

<property>
 <name>system:user.name</name>
<value>${user.name}</value>
</property>
```

-click Ctrl+s and Ctrl+x to go out the editor

#Remove log4j file:

```
$cd $HIVE_HOME/lib
```

$ ls -l log4j*.*

```
hdoop@NuvobookV1:~/apache-hive-3.1.3-bin/conf$ cd $HIVE_HOME/lib
hdoop@NuvobookV1:~/apache-hive-3.1.3-bin/lib$ ls -l log4j*.*
-rw-r--r-- 1 hdoop hdoop  208006 Jan  6  2022 log4j-1.2-api-2.17.1.jar
-rw-r--r-- 1 hdoop hdoop  301872 Jan  6  2022 log4j-api-2.17.1.jar
-rw-r--r-- 1 hdoop hdoop 1790452 Jan  6  2022 log4j-core-2.17.1.jar
-rw-r--r-- 1 hdoop hdoop   24279 Jan  6  2022 log4j-slf4j-impl-2.17.1.jar
-rw-r--r-- 1 hdoop hdoop   35962 Jan  6  2022 log4j-web-2.17.1.jar
hdoop@NuvobookV1:~/apache-hive-3.1.3-bin/lib$
```

$ rm -rf log4j-slf4j-impl-2.17.1.jar

```
hdoop@NuvobookV1:~/apache-hive-3.1.3-bin/lib$ rm -rf log4j-slf4j-impl-2.17.1
.jar
hdoop@NuvobookV1:~/apache-hive-3.1.3-bin/lib$ ls -l log4j*.*
-rw-r--r-- 1 hdoop hdoop  208006 Jan  6  2022 log4j-1.2-api-2.17.1.jar
-rw-r--r-- 1 hdoop hdoop  301872 Jan  6  2022 log4j-api-2.17.1.jar
-rw-r--r-- 1 hdoop hdoop 1790452 Jan  6  2022 log4j-core-2.17.1.jar
-rw-r--r-- 1 hdoop hdoop   35962 Jan  6  2022 log4j-web-2.17.1.jar
hdoop@NuvobookV1:~/apache-hive-3.1.3-bin/lib$
```

#insiate the local database

$ cd $HIVE_HOME/bin

$./schematool -initSchema -dbType derby

```
Initialization script completed
schemaTool completed
hdoop@NuvobookV1:~/apache-hive-3.1.3-bin/bin$
```

#Start hive

$ cd $HIVE_HOME/bin

$ ./hive

```
hdoop@hdoop-VirtualBox: ~/apache-hive-3.1.2-bin/bin
hdoop@hdoop-VirtualBox:~$ cd apache-hive-3.1.2-bin/bin
hdoop@hdoop-VirtualBox:~/apache-hive-3.1.2-bin/bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hdoop/apache-hive-3.1.2-bin/lib/log4j-sl
f4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hdoop/hadoop-3.2.1/share/hadoop/common/l
ib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = 8d24955e-52bf-43df-92b0-8f14b2f1247c

Logging initialized using configuration in jar:file:/home/hdoop/apache-hive-3.1.
2-bin/lib/hive-common-3.1.2.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versio
ns. Consider using a different execution engine (i.e. spark, tez) or using Hive
1.X releases.
Hive Session ID = 902ae3da-f1a7-4759-8c4f-cc4ef18a99c6
hive> show databases;
OK
bda
default
students
students1
students2
Time taken: 0.437 seconds, Fetched: 5 row(s)
hive>
```

>show databases;

>use default;

>show tables;

```
hdoop@hdoop-VirtualBox: ~/apache-hive-3.1.2-bin/bin
hdoop@hdoop-VirtualBox:~$ cd apache-hive-3.1.2-bin/bin
hdoop@hdoop-VirtualBox:~/apache-hive-3.1.2-bin/bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hdoop/apache-hive-3.1.2-bin/lib/log4j-sl
f4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hdoop/hadoop-3.2.1/share/hadoop/common/l
ib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = 8d24955e-52bf-43df-92b0-8f14b2f1247c

Logging initialized using configuration in jar:file:/home/hdoop/apache-hive-3.1.
2-bin/lib/hive-common-3.1.2.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versio
ns. Consider using a different execution engine (i.e. spark, tez) or using Hive
1.X releases.
Hive Session ID = 902ae3da-f1a7-4759-8c4f-cc4ef18a99c6
hive> show databases;
OK
bda
default
students
students1
students2
Time taken: 0.437 seconds, Fetched: 5 row(s)
hive>
```

Note: if hive shell command does not look like above write "exit;" to exit hiveand follow the next processor:

$ bin/hiveserver2

```
hdoop@phoenixNAP:~/apache-hive-4.0.0-bin$ bin/hiveserver2
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hdoop/apache-hive-4.0.0-bin/lib/log4j-slf4j-imp
l-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hdoop/hadoop-3.4.0/share/hadoop/common/lib/slf4
j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hdoop/apache-hive-4.0.0-bin/lib/log4j-slf4j-imp
l-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hdoop/hadoop-3.4.0/share/hadoop/common/lib/slf4
j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
2024-09-04 05:41:10: Starting HiveServer2
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hdoop/apache-hive-4.0.0-bin/lib/log4j-slf4j-imp
l-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hdoop/hadoop-3.4.0/share/hadoop/common/lib/slf4
j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = d88ad44d-0012-4809-871e-742c81e2d119
```

-Wait for the server to start and show the Hive Session ID.

- In another terminal tab, switch to the Hadoop user using the sucommand:

$ su – hdoop

$cd $HIVE_HOME

$ bin/beeline -n db_user -u jdbc:hive2://localhost:10000

```
hdoop@phoenixNAP:~/apache-hive-4.0.0-bin$ bin/beeline -n db_user -u jdbc:hive2://localh
ost:10000
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hdoop/apache-hive-4.0.0-bin/lib/log4j-slf4j-imp
l-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hdoop/hadoop-3.4.0/share/hadoop/common/lib/slf4
j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hdoop/apache-hive-4.0.0-bin/lib/log4j-slf4j-imp
l-2.18.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hdoop/hadoop-3.4.0/share/hadoop/common/lib/slf4
j-reload4j-1.7.36.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Connecting to jdbc:hive2://localhost:10000
Connected to: Apache Hive (version 4.0.0)
Driver: Hive JDBC (version 4.0.0)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 4.0.0 by Apache Hive
0: jdbc:hive2://localhost:10000>
```

>show databases;

```
0: jdbc:hive2://localhost:10000> show databases;
INFO  : Compiling command(queryId=hdoop_20240904061926_a842bdbb-180a-4342-93f1-3ddc8366
db4c): show databases
INFO  : Semantic Analysis Completed (retrial = false)
INFO  : Created Hive schema: Schema(fieldSchemas:[FieldSchema(name:database_name, type:
string, comment:from deserializer)], properties:null)
INFO  : Completed compiling command(queryId=hdoop_20240904061926_a842bdbb-180a-4342-93f
1-3ddc8366db4c); Time taken: 1.715 seconds
INFO  : Concurrency mode is disabled, not creating a lock manager
INFO  : Executing command(queryId=hdoop_20240904061926_a842bdbb-180a-4342-93f1-3ddc8366
db4c): show databases
INFO  : Starting task [Stage-0:DDL] in serial mode
INFO  : Completed executing command(queryId=hdoop_20240904061926_a842bdbb-180a-4342-93f
1-3ddc8366db4c); Time taken: 0.165 seconds
+----------------+
| database_name  |
+----------------+
| default        |
+----------------+
1 row selected (2.474 seconds)
0: jdbc:hive2://localhost:10000>
```

Install pig in linux(Ubuntu)

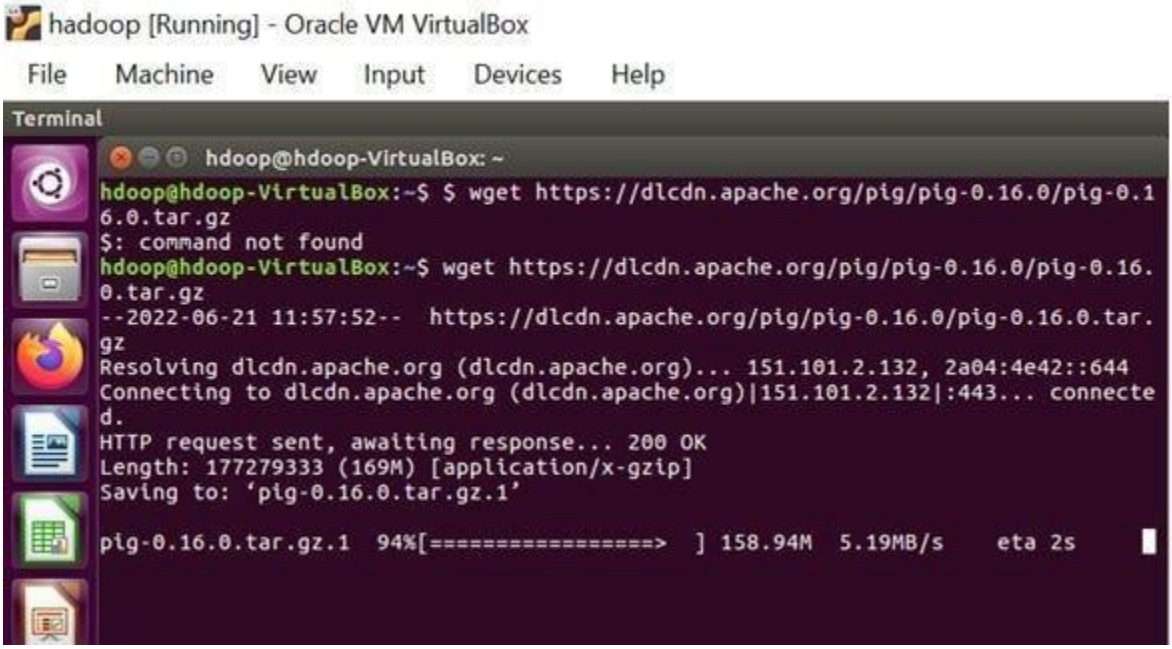
Step 1: Login into Hadoop user you used while installing Hadoop , here weuse hdoop user

```
hdoop@NuvobookV1:~$
```

$ su – hdoop

Step 2: Go to https://pig.apache.org/releases.html and copy the path of the latest version of pig that you want to install. Run the following comment to download Apache Pig in Ubuntu:

$ wget https://dlcdn.apache.org/pig/pig-0.16.0/pig-0.16.0.tar.gz



Step 3: To untar pig-0.16.0.tar.gz file run the following command:

$ tar xvzf pig-0.16.0.tar.gz

Step 4: Now open the .bashrc file to edit the path and variables/settings for pig. Run the following command:

$ sudo nano .bashrc #PIG settings

export PIG_HOME=/home/hdoop/pig-0.17.0 export PATH=$PATH:$PIG_HOME/bin

export PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_INSTALL/etc/hadoop/ export PIG_CONF_DIR=$PIG_HOME/conf

export PIG_CLASSPATH=$PIG_CONF_DIR:$PATH

```
#PIG settings
export PIG_HOME=/home/hdoop/pig-0.17.0
export PATH=$PATH:$PIG_HOME/bin
export PIG_CLASSPATH=$PIG_HOME/conf:$HADOOP_INSTALL/etc/hadoop/
export PIG_CONF_DIR=$PIG_HOME/conf
export PIG_CLASSPATH=$PIG_CONF_DIR:$PATH
```

Step 5: Run the following command to make the changes effective in the .bashrc file:

$ source .bashrc

Step 6: To start all Hadoop daemons, navigate to the hadoop-3.2.1/sbin folder and run the following commands:

$ cd hadoop-3.4.0/sbin/

$ ./start-dfs.sh

$ ./start-yarn

$ jps

```
hdoop@hdoop-VirtualBox:~$ cd hadoop-3.2.1/sbin
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [hdoop-VirtualBox]
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ jps
4817 DataNode
5298 ResourceManager
5000 SecondaryNameNode
5450 NodeManager
4683 NameNode
5982 Jps
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$
```

Step 7: Now you can launch pig by executing the following command:

$ pig

```
hadoop [Running] - Oracle VM VirtualBox
File   Machine   View   Input   Devices   Help
Terminal
hdoop@hdoop-VirtualBox: ~
hdoop@hdoop-VirtualBox:~$ pig
2022-06-21 12:29:24,863 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2022-06-21 12:29:24,864 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2022-06-21 12:29:24,864 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecT
ype
2022-06-21 12:29:24,934 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2022-06-21 12:29:24,934 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/hdoop/pig_1655794764927.log
2022-06-21 12:29:24,957 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
otup file /home/hdoop/.pigbootup not found
2022-06-21 12:29:25,273 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2022-06-21 12:29:25,273 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-06-21 12:29:25,273 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2022-06-21 12:29:25,842 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-06-21 12:29:25,862 [main] INFO  org.apache.pig.PigServer - Pig Script ID fo
r the session: PIG-default-c4e709d8-f6a3-4749-851f-c150e4603854
2022-06-21 12:29:25,862 [main] WARN  org.apache.pig.PigServer - ATS is disabled
since yarn.timeline-service.enabled set to false
grunt>
```

Step 8: Now you are in pig and can perform your desired tasks on pig. You cancome out of the pig by the

quit command:

```
> quit;
```

Program 1: Implement the following file management tasks in Hadoop:

i. **Adding files and directories**

ii. **Retrieving files**

iii. **Deleting files**

#**Adding files and directories:**

-start the Hadoop server

$ cd hadoop-3.4.0/sbin/

$ ./start-dfs.sh

$ ./start-yarn

$ jps

```
hdoop@hdoop-VirtualBox:~$ cd hadoop-3.2.1/sbin
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [hdoop-VirtualBox]
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ jps
4817 DataNode
5298 ResourceManager
5000 SecondaryNameNode
5450 NodeManager
4683 NameNode
5982 Jps
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$
```

$ sudo nano sample.txt

-write anything and click Ctrl+s and Ctrl+x to go out the editor

$hdfs dfs -mkdir /program1/

```
hdoop@NuvobookV1:~$ hdfs dfs -mkdir /program1/
```

$ hdfs dfs -copyFromLocal sample.txt /program1

```
hdoop@NuvobookV1:~$ hdfs dfs -copyFromLocal sample.txt /program1
hdoop@NuvobookV1:~$
```

$hdfs dfs -ls /program1/

```
hdoop@NuvobookV1:~$ hdfs dfs -ls /program1/
Found 1 items
-rw-r--r--   1 hdoop supergroup         24 2024-10-16 18:25 /program1/sample.txt
hdoop@NuvobookV1:~$
```

Retrieving files

$hdfs dfs -cat /program1/sample.txt

```
hdoop@NuvobookV1:~$ hdfs dfs -cat /program1/sample.txt
something written here

hdoop@NuvobookV1:~$
```

#deleting files and directory

$hdfs dfs -rm /program1/sample.txt

```
hdoop@NuvobookV1:~$ hdfs dfs -rm /program1/sample.txt
Deleted /program1/sample.txt
hdoop@NuvobookV1:~$
```

$hdfs dfs -rm -r /program1/

```
hdoop@NuvobookV1:~$ hdfs dfs -rm -r /program1/
Deleted /program1
hdoop@NuvobookV1:~$
```

Program 2: Run a basic word count Map Reduce program to understand Map Reduce Paradigm:

Login into Hadoop user you used while installing Hadoop , here we use hdoop user

```
hdoop@NuvobookV1:~$
```

$ su – hdoop

-start the Hadoop server

$ cd hadoop-3.4.0/sbin/

$ ./start-dfs.sh

$ ./start-yarn

$ jps

```
hdoop@hdoop-VirtualBox:~$ cd hadoop-3.2.1/sbin
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [hdoop-VirtualBox]
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ jps
4817 DataNode
5298 ResourceManager
5000 SecondaryNameNode
5450 NodeManager
4683 NameNode
5982 Jps
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$
```

$ sudo mkdir p2/

$ cd p2/

$ start-all.sh

$sudo nano mapper.py

-inside it paste this program:

```
#!/usr/bin/env python3

# import sys because we need to read and write data to STDIN and STDOUT
import sys

# reading entire line from STDIN (standard input)
for line in sys.stdin:
    # to remove leading and trailing whitespace
    line = line.strip()
    # split the line into words
    words = line.split()

    # we are looping over the words array and printing the word
    # with the count of 1 to the STDOUT
    for word in words:
        # write the results to STDOUT (standard output);
        # what we output here will be the input for the
```

```
        # Reduce step, i.e. the input for reducer.py
        print ('%s\t%s' % (word, 1))
```

-click Ctrl+s and Ctrl+x to close it

$ nano reducer.py

-insert this program inside it:

```
#!/usr/bin/env python3

from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None

# read the entire line from STDIN
for line in sys.stdin:
    # remove leading and trailing whitespace
    line = line.strip()
    # splitting the data on the basis of tab we have provided in mapper.py
```

```
    word, count = line.split('\t', 1)
    # convert count (currently a string) to int
    try:
        count = int(count)
    except ValueError:
        # count was not a number, so silently
        # ignore/discard this line
        continue

    # this IF-switch only works because Hadoop sorts map output
    # by key (here: word) before it is passed to the reducer
    if current_word == word:
        current_count += count
    else:
        if current_word:
            # write result to STDOUT
            print ('%s\t%s' % (current_word, current_count))
        current_count = count
        current_word = word

# do not forget to output the last word if needed!
if current_word == word:
    print ('%s\t%s' % (current_word, current_count))
```

-click Ctrl+s and Ctrl+x to close it

$ sudo sample.txt

-insert this program inside it:

The server processes data, repeating tasks to repeat actions efficiently. Users rely on it, repeating operations to repeat access to critical data

-click Ctrl+s and Ctrl+x to close it

$ sudo chmod 777 reducer.py

$sudo chmod 777 mapper.py

$ hdfs dfs -mkdir /p2/

$ hdfs dfs -copyFromLocal sample.txt /p2/

$ cd

$ wget https://repo1.maven.org/maven2/org/apache/hadoop/hadoop-streaming/2.7.3/hadoop-streaming-2.7.3.jar

$ hadoop jar /home/hdoop/hadoop-streaming-2.7.3.jar -input /p2/sample.txt -output /p2/output -mapper /home/hdoop/p2/mapper.py -reducer /home/hdoop/p2/reducer.py

#Output:

```
hdoop@NuvobookV1:~/lab/p2$ hadoop jar /home/hdoop/hadoop-streaming-2.7.3.jar -input /p2/python/sample.txt -output /p2/python/output -mapper /home/hdoop/lab/
p2/mapper.py -reducer /home/hdoop/lab/p2/reducer.py
packageJobJar: [/tmp/hadoop-unjar4769045384438272080/] [] /tmp/streamjob7249487522430462 05.jar tmpDir=null
2024-10-16 17:08:52,535 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /172.18.132.51:8032
2024-10-16 17:08:52,799 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /172.18.132.51:8032
2024-10-16 17:08:53,145 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_1729063021124_001
1
2024-10-16 17:08:53,409 INFO mapred.FileInputFormat: Total input files to process : 1
2024-10-16 17:08:53,489 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-16 17:08:53,674 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1729063021124_0011
2024-10-16 17:08:53,674 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-16 17:08:53,890 INFO conf.Configuration: resource-types.xml not found
2024-10-16 17:08:53,890 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-16 17:08:54,019 INFO impl.YarnClientImpl: Submitted application application_1729063021124_0011
2024-10-16 17:08:54,062 INFO mapreduce.Job: The url to track the job: http://172.18.132.51:8088/proxy/application_1729063021124_0011/
2024-10-16 17:08:54,064 INFO mapreduce.Job: Running job: job_1729063021124_0011
2024-10-16 17:09:01,400 INFO mapreduce.Job: Job job_1729063021124_0011 running in uber mode : false
2024-10-16 17:09:01,402 INFO mapreduce.Job:  map 0% reduce 0%
2024-10-16 17:09:06,470 INFO mapreduce.Job:  map 100% reduce 0%
2024-10-16 17:09:11,508 INFO mapreduce.Job:  map 100% reduce 100%
2024-10-16 17:09:13,541 INFO mapreduce.Job: Job job_1729063021124_0011 completed successfully
2024-10-16 17:09:13,648 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=241
                FILE: Number of bytes written=934444
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=417
                HDFS: Number of bytes written=160
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=5124
                Total time spent by all reduces in occupied slots (ms)=2764
                Total time spent by all map tasks (ms)=5124
                Total time spent by all reduce tasks (ms)=2764
                Total vcore-milliseconds taken by all map tasks=5124
```

```
                Total time spent by all reduce tasks (ms)=2764
                Total vcore-milliseconds taken by all map tasks=5124
                Total vcore-milliseconds taken by all reduce tasks=2764
                Total megabyte-milliseconds taken by all map tasks=5246976
                Total megabyte-milliseconds taken by all reduce tasks=2830336
        Map-Reduce Framework
                Map input records=1
                Map output records=22
                Map output bytes=191
                Map output materialized bytes=247
                Input split bytes=196
                Combine input records=0
                Combine output records=0
                Reduce input groups=18
                Reduce shuffle bytes=247
                Reduce input records=22
                Reduce output records=18
                Spilled Records=44
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=162
                CPU time spent (ms)=2180
                Physical memory (bytes) snapshot=885735424
                Virtual memory (bytes) snapshot=7685988352
                Total committed heap usage (bytes)=678952960
                Peak Map Physical memory (bytes)=323989504
                Peak Map Virtual memory (bytes)=2561187840
                Peak Reduce Physical memory (bytes)=240586752
                Peak Reduce Virtual memory (bytes)=2564431872
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=221
        File Output Format Counters
                Bytes Written=160
2024-10-16 17:09:13,649 INFO streaming.StreamJob: Output directory: /p2/python/output
hdoop@NuvobookV1:~/lab/p2$
```

Program 3: Write a Map Reduce program that mines weather data. Hint: Weather sensors collecting data every hour at many locations across the globe gather a large volume of log data, which is a good candidate for analysis with Map Reduce, since it is semi structured and record oriented.

: Login into Hadoop user you used while installing Hadoop , here we use hdoop user

```
hdoop@NuvobookV1:~$
```

$ su – hdoop

-start the Hadoop server

$ cd hadoop-3.4.0/sbin/

$ ./start-dfs.sh

$ ./start-yarn

$ jps

```
hdoop@hdoop-VirtualBox:~$ cd hadoop-3.2.1/sbin
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [hdoop-VirtualBox]
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ jps
4817 DataNode
5298 ResourceManager
5000 SecondaryNameNode
5450 NodeManager
4683 NameNode
5982 Jps
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$
```

#make a dir for the program 3

$ hdfs dfs -mkdir /p3

```
hdoop@NuvobookV1:~$ hdfs dfs -mkdir /p3
hdoop@NuvobookV1:~$
```

#create input and output folders

$ hdfs dfs -mkdir /p3/input

$ hdfs dfs -mkdir /p3/output

```
hdoop@NuvobookV1:~$ hdfs dfs -ls /p3/
Found 2 items
drwxr-xr-x   - hdoop supergroup          0 2024-10-16 18:42 /p3/input
drwxr-xr-x   - hdoop supergroup          0 2024-10-16 18:42 /p3/output
hdoop@NuvobookV1:~$
```

#create mapper program

$sudo mkdir program3/

$cd program3/

$sudo nano mapper.py

-insert this code

```
#!/usr/bin/env python3


import sys


# input comes from STDIN (standard input)
# the mapper will get daily max temperature and group it by month. so output will be (month,dailymax_temperature)
for line in sys.stdin:
    # remove leading and trailing whitespace
```

```
    line = line.strip()

    # split the line into words

    words = line.split()

    #See the README hosted on the weather website  which help us understand how each position represents a column

    month = line[10:12]

    daily_max = line[38:45]

    daily_max = daily_max.strip()

    # increase counters

    for word in words:

        # write the results to STDOUT (standard output);

        # what we output here will be go through the shuffle proess and then


        # be the input for the Reduce step, i.e. the input for reducer.py

        #

        # tab-delimited; month and daily max temperature as output

        print ('%s\t%s' % (month ,daily_max))
```

-click Ctrl+s and Ctrl+x to close it

note: make sure you have python3 in your system

#reducer program

$sudo nano reducer.py

-insert this code

```
#!/usr/bin/env python3


from operator import itemgetter

import sys

#reducer will get the input from stdid which will be a collection of key, value(Key=month , value= daily max temperature)

#reducer logic: will get all the daily max temperature for a month and find max temperature for the month

#shuffle will ensure that key are sorted(month)

current_month = None

current_max = 0

month = None


# input comes from STDIN

for line in sys.stdin:

    # remove leading and trailing whitespace

    line = line.strip()

    # parse the input we got from mapper.py

    month, daily_max = line.split('\t', 1)


    # convert daily_max (currently a string) to float
```

```
    try:
        daily_max = float(daily_max)
    except ValueError:
        # daily_max was not a number, so silently
        # ignore/discard this line
        continue

    # this IF-switch only works because Hadoop shuffle process sorts map output
    # by key (here: month) before it is passed to the reducer
    if current_month == month:
        if daily_max > current_max:
            current_max = daily_max
    else:
        if current_month:
            # write result to STDOUT
            print ('%s\t%s' % (current_month, current_max))
        current_max = daily_max
        current_month = month

# output of the last month
if current_month == month:
    print ('%s\t%s' % (current_month, current_max))
```

#create the dataset of temp :

Copy the content of this web page:

https://www.ncei.noaa.gov/pub/data/uscrn/products/daily01/2002/CRND0103-2002-RI_Kingston_1_NW.txt

$ sudo nano tempdata.txt

-insert the data u got from the webpage

#change the permissions of the files

$sudo chmod 777 mapper.py reducer.py

#upload the dataset to Hadoop

$ hdfs dfs -copyFromLocal tempdatanew.txt /p3/input

```
hdoop@NuvobookV1:~/lab/p3$ hdfs dfs -copyFromLocal tempdatanew.txt /p3/input
hdoop@NuvobookV1:~/lab/p3$ hdfs dfs -ls /p3/input
Found 1 items
-rw-r--r--   1 hdoop supergroup      79205 2024-10-16 18:58 /p3/input/tempda
tanew.txt
hdoop@NuvobookV1:~/lab/p3$
```

#running python program in Hadoop using Hadoop streamer

$ cd

$ wget https://repo1.maven.org/maven2/org/apache/hadoop/hadoopstreaming/2.7.3/hadoop-streaming-2.7.3.jar $ hadoop jar /home/hdoop/hadoop-streaming-2.7.3.jar -input /p2/sample.txt -output /p2/output -mapper /home/hdoop/p2/mapper.py -reducer /home/hdoop/p2/reducer.py

$ hadoop jar /home/hdoop/hadoop-streaming-2.7.3.jar -input /p3/input/tempdatanew.txt -output /p3/output/outputUpdate -mapper /home/hdoop/p3/mapper.py -reducer /home/hdoop/p3/reducer.py

#output

```
hdoop@NuvobookV1:~/lab/p3$ hadoop jar /home/hdoop/hadoop-streaming-2.7.3.jar  -input /p3/input/tempdatanew.txt -output /p3/output/outputUpdate -mapper /home
/hdoop/lab/p3/mapper.py -reducer /home/hdoop/lab/p3/reducer.py
packageJobJar: [/tmp/hadoop-unjar7804424084048329055/] [] /tmp/streamjob5131961797712801057.jar tmpDir=null
2024-10-16 19:11:32,960 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /172.18.132.51:8032
2024-10-16 19:11:33,189 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /172.18.132.51:8032
2024-10-16 19:11:33,508 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_1729063021124_001
3
2024-10-16 19:11:33,835 INFO mapred.FileInputFormat: Total input files to process : 1
2024-10-16 19:11:34,308 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-16 19:11:34,929 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1729063021124_0013
2024-10-16 19:11:34,929 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-16 19:11:35,149 INFO conf.Configuration: resource-types.xml not found
2024-10-16 19:11:35,150 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-16 19:11:35,627 INFO impl.YarnClientImpl: Submitted application application_1729063021124_0013
2024-10-16 19:11:35,683 INFO mapreduce.Job: The url to track the job: http://172.18.132.51:8088/proxy/application_1729063021124_0013/
2024-10-16 19:11:35,685 INFO mapreduce.Job: Running job: job_1729063021124_0013
2024-10-16 19:11:41,855 INFO mapreduce.Job: Job job_1729063021124_0013 running in uber mode : false
2024-10-16 19:11:41,857 INFO mapreduce.Job:  map 0% reduce 0%
2024-10-16 19:11:47,950 INFO mapreduce.Job:  map 100% reduce 0%
2024-10-16 19:11:54,828 INFO mapreduce.Job:  map 100% reduce 100%
2024-10-16 19:11:55,854 INFO mapreduce.Job: Job job_1729063021124_0013 completed successfully
2024-10-16 19:11:55,989 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=103494
                FILE: Number of bytes written=1140983
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=83505
                HDFS: Number of bytes written=96
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=6156
                Total time spent by all reduces in occupied slots (ms)=4504
                Total time spent by all map tasks (ms)=6156
                Total time spent by all reduce tasks (ms)=4504
                Total vcore-milliseconds taken by all map tasks=6156
                Total time spent by all reduces in occupied slots (ms)=4504
                Total time spent by all map tasks (ms)=6156
                Total time spent by all reduce tasks (ms)=4504
                Total vcore-milliseconds taken by all map tasks=6156
                Total vcore-milliseconds taken by all reduce tasks=4504
                Total megabyte-milliseconds taken by all map tasks=6303744
                Total megabyte-milliseconds taken by all reduce tasks=4612096
        Map-Reduce Framework
                Map input records=365
                Map output records=10220
                Map output bytes=83048
                Map output materialized bytes=103500
                Input split bytes=204
                Combine input records=0
                Combine output records=0
                Reduce input groups=12
                Reduce shuffle bytes=103500
                Reduce input records=10220
                Reduce output records=12
                Spilled Records=20440
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=172
                CPU time spent (ms)=3120
                Physical memory (bytes) snapshot=898703360
                Virtual memory (bytes) snapshot=7689854976
                Total committed heap usage (bytes)=675282944
                Peak Map Physical memory (bytes)=328138752
                Peak Map Virtual memory (bytes)=2562195456
                Peak Reduce Physical memory (bytes)=244801536
                Peak Reduce Virtual memory (bytes)=2567061504
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=83301
        File Output Format Counters
                Bytes Written=96
```

Program 4: Implement matrix multiplication with Hadoop Map Reduce.

: Login into Hadoop user you used while installing Hadoop , here we use hdoop user

```
hdoop@NuvobookV1:~$
```

$ su – hdoop

-start the Hadoop server

$ cd hadoop-3.4.0/sbin/

$ ./start-dfs.sh

$ ./start-yarn

$ jps

```
hdoop@hdoop-VirtualBox:~$ cd hadoop-3.2.1/sbin
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [hdoop-VirtualBox]
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ jps
4817 DataNode
5298 ResourceManager
5000 SecondaryNameNode
5450 NodeManager
4683 NameNode
5982 Jps
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$
```

#make a dir for the program 4

$ hdfs dfs -mkdir /p4

```
ndoop@NuvobookV1:~/lab/p4$ hdfs dfs -mkdir /p4/
ndoop@NuvobookV1:~/lab/p4$
```

#create mapper program

$sudo nano mapper.py

-insert this code

```
#!/usr/bin/env python3


import sys


for line in sys.stdin:
    line = line.strip()
    entry = line.split(',')
    key=entry[0]
    value = line[1:]
    if key =='A':
        print('{0}\t{1}'.format(key,value))
    elif key =='B':
        print('{0}\t{1}'.format(key,value))
```

-click Ctrl+s and Ctrl+x to close it

note: make sure you have python3 in your system

#reducer program

$sudo nano reducer.py

-insert this code

```
#!/usr/bin/env python

import sys

mat1 = {}

mat2 = {}


# Reading input and populating the matrices

for line in sys.stdin:

    line = line.strip()

    key, value = line.split('\t')

    v = value.split(',')

    if key == 'A':

        mat1[(int(v[1]), int(v[2]))] = int(v[3])

    elif key == 'B':  # Corrected from '8' to 'B'

        mat2[(int(v[1]), int(v[2]))] = int(v[3])


result = 0

# Multiplying the matrices with safe access to the elements

for i in range(0, 3):

    for j in range(0, 3):

        for k in range(0, 3):
```

```
        # Using .get() method to avoid KeyError
        result += mat1.get((i, k), 0) * mat2.get((k, j), 0)
    print('({0},{1})\t{2}'.format(i, j, result))
    result = 0
```

#create the matrix file :

$ sudo nano matrixinput.txt

-insert the data

```
A,0,0,1
A,0,1,2
A,0,2,3
A,1,0,4
A,1,1,5
A,1,2,6
A,2,0,7
A,2,1,8
A,2,2,9

B,0,0,1
B,0,1,1
B,0,2,1
B,1,0,1
B,1,1,2
```

B,1,2,3

B,2,0,1

B,2,1,1

B,2,2,1

#change the permissions of the files

$sudo chmod 777 mapper.py reducer.py

#upload the dataset to Hadoop

$ hdfs dfs -copyFromLocal matrixinput.txt /p4/

```
hdoop@NuvobookV1:~/lab/p4$ hdfs dfs -copyFromLocal matrixinput.txt /p4/
hdoop@NuvobookV1:~/lab/p4$ hdfs dfs -ls /p4
Found 1 items
-rw-r--r--   1 hdoop supergroup        145 2024-10-17 07:12 /p4/matrixinput.
txt
hdoop@NuvobookV1:~/lab/p4$
```

#running python program in Hadoop using Hadoop streamer

$ cd

$ wget https://repo1.maven.org/maven2/org/apache/hadoop/hadoopstreaming/2.7.3/hadoop-streaming-2.7.3.jar $ hadoop jar /home/hdoop/hadoop-streaming-2.7.3.jar -input /p2/sample.txt -output /p2/output -mapper /home/hdoop/p2/mapper.py -reducer /home/hdoop/p2/reducer.py

$ hadoop jar /home/hdoop/hadoop-streaming-2.7.3.jar -input /p4/matrixinput.txt -output /p4/output -mapper /home/hdoop/mapper.py -reducer /home/hdoop/reducer.py

#output

```
hdoop@NuvobookV1:~/lab/p4$ hadoop jar /home/hdoop/hadoop-streaming-2.7.3.jar -input /p4/matrixinput.txt -output /p4/output -mapper /home/hdoop/lab/p4/mapper
.py -reducer /home/hdoop/lab/p4/reducer.py
packageJobJar: [/tmp/hadoop-unjar6029587034847901962/] [] /tmp/streamjob6834679690450673703.jar tmpDir=null
2024-10-17 07:13:55,545 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /172.18.132.51:8032
2024-10-17 07:13:55,839 INFO client.DefaultNoHARMFailoverProxyProvider: Connecting to ResourceManager at /172.18.132.51:8032
2024-10-17 07:13:56,221 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_1729128507769_000
9
2024-10-17 07:13:56,504 INFO mapred.FileInputFormat: Total input files to process : 1
2024-10-17 07:13:56,585 INFO mapreduce.JobSubmitter: number of splits:2
2024-10-17 07:13:56,788 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1729128507769_0009
2024-10-17 07:13:56,789 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-10-17 07:13:57,003 INFO conf.Configuration: resource-types.xml not found
2024-10-17 07:13:57,004 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2024-10-17 07:13:57,431 INFO impl.YarnClientImpl: Submitted application application_1729128507769_0009
2024-10-17 07:13:57,471 INFO mapreduce.Job: The url to track the job: http://172.18.132.51:8088/proxy/application_1729128507769_0009/
2024-10-17 07:13:57,473 INFO mapreduce.Job: Running job: job_1729128507769_0009
2024-10-17 07:14:04,625 INFO mapreduce.Job: Job job_1729128507769_0009 running in uber mode : false
2024-10-17 07:14:04,626 INFO mapreduce.Job:  map 0% reduce 0%
2024-10-17 07:14:11,870 INFO mapreduce.Job:  map 100% reduce 0%
2024-10-17 07:14:16,917 INFO mapreduce.Job:  map 100% reduce 100%
2024-10-17 07:14:18,955 INFO mapreduce.Job: Job job_1729128507769_0009 completed successfully
2024-10-17 07:14:19,192 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=204
                FILE: Number of bytes written=934346
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=410
                HDFS: Number of bytes written=79
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots (ms)=9464
                Total time spent by all reduces in occupied slots (ms)=2657
                Total time spent by all map tasks (ms)=9464
                Total time spent by all reduce tasks (ms)=2657
                Total vcore-milliseconds taken by all map tasks=9464
                Total vcore-milliseconds taken by all reduce tasks=2657
```

```
                Total vcore-milliseconds taken by all map tasks=9464
                Total vcore-milliseconds taken by all reduce tasks=2657
                Total megabyte-milliseconds taken by all map tasks=9691136
                Total megabyte-milliseconds taken by all reduce tasks=2720768
        Map-Reduce Framework
                Map input records=19
                Map output records=18
                Map output bytes=162
                Map output materialized bytes=210
                Input split bytes=192
                Combine input records=0
                Combine output records=0
                Reduce input groups=2
                Reduce shuffle bytes=210
                Reduce input records=18
                Reduce output records=9
                Spilled Records=36
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=183
                CPU time spent (ms)=3290
                Physical memory (bytes) snapshot=878444544
                Virtual memory (bytes) snapshot=7690653696
                Total committed heap usage (bytes)=668991488
                Peak Map Physical memory (bytes)=320208896
                Peak Map Virtual memory (bytes)=2560774144
                Peak Reduce Physical memory (bytes)=238653440
                Peak Reduce Virtual memory (bytes)=2570227712
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=218
        File Output Format Counters
                Bytes Written=79
2024-10-17 07:14:19,192 INFO streaming.StreamJob: Output directory: /p4/output
hdoop@NuvobookV1:~/lab/p4$
```

$ hdfs dfs -cat /p4/output/part-00000

```
hdoop@NuvobookV1:~/lab/p4$ hdfs dfs -cat /p4/output/part-00000
(0,0)   6
(0,1)   8
(0,2)   10
(1,0)   15
(1,1)   20
(1,2)   25
(2,0)   24
(2,1)   32
(2,2)   40
hdoop@NuvobookV1:~/lab/p4$
```

Program 5: Run the Pig Latin Scripts to find Word Count.

Login into Hadoop user you used while installing Hadoop , here we use hdoop user

```
hdoop@NuvobookV1:~$
```

$ su – hdoop

-start the Hadoop server

$ cd hadoop-3.4.0/sbin/

$ ./start-dfs.sh

$ ./start-yarn

$ jps

```
hdoop@hdoop-VirtualBox:~$ cd hadoop-3.2.1/sbin
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [hdoop-VirtualBox]
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ jps
4817 DataNode
5298 ResourceManager
5000 SecondaryNameNode
5450 NodeManager
4683 NameNode
5982 Jps
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$
```

$ sudo mkdir p5/

$ cd p5/

$sudo nano countname.txt

-Insert this:

Alice, Bob, Charlie, Alice, David, Eve, Frank, Charlie, Bob, Grace, Alice, Frank, David, Helen, Eve, Frank, Bob

```
hdoop@NuvobookV1:~/lab/p5$ sudo nano countname.txt
[sudo] password for hdoop:
hdoop@NuvobookV1:~/lab/p5$ cat countname.txt
Alice, Bob, Charlie, Alice, David, Eve, Frank, Charlie, Bob, Grace, Alice, F
rank, David, Helen, Eve, Frank, Bob.
```

#upload the file to Hadoop:

$ hdfs dfs -mkdir /p5/

$ hdfs dfs -mkdir /p5/input

$ hdfs dfs -copyFromLocal /home/hdoop/p5/countname.txt /p5/input

```
hdoop@NuvobookV1:~/lab/p5$ hdfs dfs -ls /p5/
Found 1 items
drwxrwxr-x   - hdoop supergroup          0 2024-10-16 19:58 /p5/input
hdoop@NuvobookV1:~/lab/p5$ hdfs dfs -ls /p5/input
Found 1 items
-rw-r--r--   1 hdoop supergroup        113 2024-10-16 19:58 /p5/input/countname.txt
```

#run pig :

$ pig -x mapreduce

```
hdoop@NuvobookV1:~/lab/p5$ pig -x mapreduce
2024-10-16 19:59:03,530 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-16 19:59:03,532 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDU
CE
2024-10-16 19:59:03,532 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the E
xecType
2024-10-16 19:59:03,588 WARN pig.Main: Cannot write to log file: /home/hdoop
/lab/p5/pig_1729108743588.log
2024-10-16 19:59:03,599 [main] INFO  org.apache.pig.Main - Apache Pig versio
n 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-10-16 19:59:03,635 [main] INFO  org.apache.pig.impl.util.Utils - Defaul
t bootup file /home/hdoop/.pigbootup not found
2024-10-16 19:59:03,918 [main] INFO  org.apache.hadoop.conf.Configuration.de
precation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtrac
ker.address
2024-10-16 19:59:03,918 [main] INFO  org.apache.pig.backend.hadoop.execution
engine.HExecutionEngine - Connecting to hadoop file system at: hdfs://172.18
.132.51:9000
2024-10-16 19:59:04,458 [main] INFO  org.apache.pig.PigServer - Pig Script I
D for the session: PIG-default-7825e024-6755-4975-aad7-14900747aff5
2024-10-16 19:59:04,458 [main] WARN  org.apache.pig.PigServer - ATS is disab
led since yarn.timeline-service.enabled set to false
grunt> |
```

#write map reduce program in pig:

-load the file countname that we upload to Hadoop:

input_lines = LOAD '/p5/input/countname.txt' AS (line:chararray);

```
grunt> input_lines = LOAD '/p5/input/countname.txt' AS (line:chararray);
```

-the rest of program which it will count the names:

Note: insert each line separately ,it should end up with “;”

words = FOREACH input_lines GENERATE FLATTEN(TOKENIZE(line)) AS word;

filtered_words = FILTER words BY word MATCHES ‘\\w+’;

word_groups = GROUP filtered_words BY word;

word_count = FOREACH word_groups GENERATE COUNT(filtered_words) AS count, group AS word;

ordered_word_count = ORDER word_count BY count DESC;
DUMP ordered_word_count;

```
grunt> filtered_words = FILTER words BY word MATCHES '\\w+';
grunt> word_groups = GROUP filtered_words BY word;
grunt> word_count = FOREACH word_groups GENERATE COUNT(filtered_words) AS co
unt, group AS word;
grunt> ordered_word_count = ORDER word_count BY count DESC;
grunt> DUMP ordered_word_count;
```

Note: it will take approx. 10 minutes to compile.

#Output:

```
2024-10-16 20:18:36,993 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths t
o process : 1
(3,Frank)
(3,Alice)
(2,Charlie)
(2,David)
(2,Eve)
(2,Bob)
(1,Helen)
(1,Grace)
grunt> quit
2024-10-16 20:22:41,830 [main] INFO  org.apache.pig.Main - Pig script completed in 23 minutes, 38 seconds and 336 millis
econds (1418336 ms)
hdoop@NuvobookV1:~/lab/p5$ hdfs dfs -rm -r /p5/output
```

-write quit to stop pig

#stop Hadoop:

$stop-all.sh

Program 6: Construct the Pig Latin Scripts to find a max temp for each and every year.

Login into Hadoop user you used while installing Hadoop , here we use hdoop user

```
hdoop@NuvobookV1:~$
```

$ su – hdoop

-start the Hadoop server

$ cd hadoop-3.4.0/sbin/
$ ./start-dfs.sh
$ ./start-yarn
$ jps

```
hdoop@hdoop-VirtualBox:~$ cd hadoop-3.2.1/sbin
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [hdoop-VirtualBox]
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$ jps
4817 DataNode
5298 ResourceManager
5000 SecondaryNameNode
5450 NodeManager
4683 NameNode
5982 Jps
hdoop@hdoop-VirtualBox:~/hadoop-3.2.1/sbin$
```

$ sudo mkdir p6/
$ cd p5/
#create the dataset:
$sudo nano temperature.txt

-Insert this dataset:

1938 5.0
1938 13.3
1941 33.9
1974 8.9
1974 8.9
1983 31.7
1983 30.6
1991 13.3
1993 32.8
1950 7.2
1951 8.3
1955 10.0
1960 19.4

1961	22.2
1965	17.2
1966	23.9
1970	25.6
1972	27.8
1975	27.2
1976	22.8
1980	18.3
1981	31.1
1984	28.3
1985	28.9
1990	21.1
1992	29.4
1994	32.2
1995	20.0
1996	15.6
2000	18.9
2001	26.7
2002	25.0
2003	26.1
2005	30.0
2006	23.9
2008	27.8
2010	27.2
2012	20.6
2013	31.1
2014	33.3
2015	30.6
2016	33.9
2017	22.2
2018	32.8
2019	31.7
2020	34.4
2021	31.1
2022	33.3
2023	35.0
2024	29.4

```
hdoop@NuvobookV1:~$ cd lab/p6/
hdoop@NuvobookV1:~/lab/p6$ cat temperature.txt
1938 5.0
1938 13.3
1941 33.9
1974 8.9
1974 8.9
1983 31.7
1983 30.6
1991 13.3
1993 32.8
1950 7.2
1951 8.3
1955 10.0
1960 19.4
1961 22.2
1965 17.2
1966 23.9
1970 25.6
1972 27.8
1975 27.2
1976 22.8
1980 18.3
1981 31.1
1984 28.3
1985 28.9
1990 21.1
1992 29.4
1994 32.2
1995 20.0
1996 15.6
2000 18.9
2001 26.7
2002 25.0
2003 26.1
2005 30.0
2006 23.9
2008 27.8
2010 27.2
```

#upload the file to Hadoop:
$ hdfs dfs -mkdir /p6/
$hadoop fs -put temperature.txt /p6/

```
hdoop@NuvobookV1:~/lab/p6$ hadoop fs -ls /p6
Found 1 items
-rw-r--r--   1 hdoop supergroup        495 2024-10-18 03:01 /p6/temper
ature.txt
hdoop@NuvobookV1:~/lab/p6$
```

#run pig :

$ pig

```
hdoop@NuvobookV1:~/lab/p5$ pig -x mapreduce
2024-10-16 19:59:03,530 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2024-10-16 19:59:03,532 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDU
CE
2024-10-16 19:59:03,532 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the E
xecType
2024-10-16 19:59:03,588 WARN pig.Main: Cannot write to log file: /home/hdoop
/lab/p5/pig_1729108743588.log
2024-10-16 19:59:03,599 [main] INFO  org.apache.pig.Main - Apache Pig versio
n 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2024-10-16 19:59:03,635 [main] INFO  org.apache.pig.impl.util.Utils - Defaul
t bootup file /home/hdoop/.pigbootup not found
2024-10-16 19:59:03,918 [main] INFO  org.apache.hadoop.conf.Configuration.de
precation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtrac
ker.address
2024-10-16 19:59:03,918 [main] INFO  org.apache.pig.backend.hadoop.execution
engine.HExecutionEngine - Connecting to hadoop file system at: hdfs://172.18
.132.51:9000
2024-10-16 19:59:04,458 [main] INFO  org.apache.pig.PigServer - Pig Script I
D for the session: PIG-default-7825e024-6755-4975-aad7-14900747aff5
2024-10-16 19:59:04,458 [main] WARN  org.apache.pig.PigServer - ATS is disab
led since yarn.timeline-service.enabled set to false
grunt> |
```

#write this query in pig:

-load the file temperature that we upload to Hadoop:

data = LOAD '/p6/temperature.txt' USING PigStorage(' ') AS (year:int, temp:int);

```
grunt> data = LOAD '/p6/temperature.txt' USING PigStorage(' ') AS (year:int,
 temp:int);
2024-10-18 03:15:00,345 [main] INFO  org.apache.hadoop.conf.Configuration.de
precation - yarn.resourcemanager.system-metrics-publisher.enabled is depreca
ted. Instead, use yarn.system-metrics-publisher.enabled
grunt> |
```

- Group the data by year:

grouped_data = GROUP data BY year;

- Find the maximum temperature for each year

max_temp_by_year = FOREACH grouped_data GENERATE group AS year, MAX(data.temp) AS max_temp;

- Store the result in an output file

STORE max_temp_by_year INTO '/p6/max_temperature_per_year' USING PigStorage(',');

```
grunt> data = LOAD '/p6/temperature.txt' USING PigStorage(' ') AS (year:int,
 temp:int);
2024-10-18 03:15:00,345 [main] INFO  org.apache.hadoop.conf.Configuration.de
precation - yarn.resourcemanager.system-metrics-publisher.enabled is depreca
ted. Instead, use yarn.system-metrics-publisher.enabled
grunt> grouped_data = GROUP data BY year;
grunt> max_temp_by_year = FOREACH grouped_data GENERATE group AS year, MAX(d
ata.temp) AS max_temp;
grunt> STORE max_temp_by_year INTO '/p6/max_temperature_per_year' USING PigS
torage(',');
```

Note: it will take approx. 10 minutes to compile.

```
engine.mapReduceLayer.MapReduceLauncher - Success!
grunt>
```

Output:

-in new terminal log in to hdoop again and chechk the output by writing this command line (it print out the content of your out put file that is inside Hadoop file system)

$hadoop fs -cat /p6/max_temperature_per_year/part-r-00000

```
hdoop@NuvobookV1:~/lab/p6$ hadoop fs -cat /p6/max_temperature_per_year
/part-r-00000
1938,13
1941,33
1950,7
1951,8
1955,10
1960,19
1961,22
1965,17
1966,23
1970,25
1972,27
1974,8
1975,27
1976,22
1980,18
1981,31
1983,31
1984,28
1985,28
1990,21
1991,13
1992,29
1993,32
1994,32
1995,20
1996,15
2000,18
2001,26
2002,25
2003,26
2005,30
2006,23
2008,27
2010,27
```

Note:” part-r-00000” file name might change , to check the exact name of the file run this line :

$ hadoop fs -ls /p6/max_temperature_per_year

```
hdoop@NuvobookV1:~/lab/p6$ hadoop fs -ls /p6/max_temperature_per_year
Found 2 items
-rw-r--r--   1 hdoop supergroup          0 2024-10-18 03:19 /p6/max_te
mperature_per_year/_SUCCESS
-rw-r--r--   1 hdoop supergroup        373 2024-10-18 03:19 /p6/max_te
mperature_per_year/part-r-00000
```

-write quit to stop pig

```
grunt> quit
2024-10-18 03:27:31,573 [main] INFO  org.apache.pig.Main - Pig script comple
ted in 54 minutes, 8 seconds and 763 milliseconds (3248763 ms)
hdoop@NuvobookV1:~$
```

#stop Hadoop:
$stop-all.sh

Program 7: Use Hive to create, alter, and drop databases, tables, views, functions, and indexes.

#create and show database:

hive> CREATE SCHEMA userdb;

hive> SHOW DATABASES;

```
hive> show tables;
OK
Time taken: 0.71 seconds
hive> CREATE SCHEMA userdb;
OK
Time taken: 0.28 seconds
hive> SHOW DATABASES;
OK
default
userdb
Time taken: 0.038 seconds, Fetched: 2 row(s)
hive>
```

#Drop Database Statement

hive> DROP SCHEMA userdb;

```
hive> DROP SCHEMA userdb;
OK
Time taken: 0.316 seconds
hive>
```

#Create and show Table Statement

hive>CREATE TABLE employee (eid int, name String, salary String, destination

String)COMMENT 'Employee details' ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE;

```
hive> CREATE TABLE employee (eid int, name String, salary String, destination String)COMMENT 'Employee details' ROW FORMAT DELIMITED FI
ELDS TERMINATED BY ',' LINES TERMINATED BY '\n' STORED AS TEXTFILE;
OK
Time taken: 0.729 seconds
hive>
```

hive> show tables

```
hive> show tables;
OK
employee
Time taken: 0.035 seconds, Fetched: 1 row(s)
hive>
```

hive> describe employee;

```
hive> describe employee;
OK
eid                     int
name                    string
salary                  string
destination             string
Time taken: 0.067 seconds, Fetched: 4 row(s)
hive>
```

#run quary by laoding from a local file :

hive> LOAD DATA LOCAL INPATH '/home/hdoop/sample.txt' OVERWRITE INTO TABLE employee;

```
hive> LOAD DATA LOCAL INPATH '/home/hdoop/sample.txt' OVERWRITE INTO TABLE employee;
Loading data to table default.employee
OK
Time taken: 1.588 seconds
hive>
```

#get all rows from a table

hive> SELECT * FROM employee;

```
hive> SELECT * FROM employee;
OK
NULL    "Gopal" "45000" Technical"manager"
NULL    "Manisha"       "45000" "Proof reader"
NULL    "Masthanvali"   "40000" "Technical writer"
NULL    "Kiran" "40000" "Hr Admin"
NULL    "Kranthi"       "30000" "Op Admin"
Time taken: 2.036 seconds, Fetched: 5 row(s)
hive>
```

#modify table name:

hive> ALTER TABLE employee RENAME TO emp;

```
hive> ALTER TABLE employee RENAME TO emp;
OK
Time taken: 0.249 seconds
hive> show tables;
OK
emp
Time taken: 0.043 seconds, Fetched: 1 row(s)
hive>
```

#Drop Table Statement

```
hive> DROP TABLE IF EXISTS emp;
```

```
hive> DROP TABLE IF EXISTS emp;
OK
Time taken: 0.567 seconds
hive>
```

#Creating a View

```
hive> CREATE VIEW emp_30000 AS
SELECT * FROM employee
WHERE salary>30000;
```

```
hive> CREATE VIEW emp_30000 AS
    > SELECT * FROM employee
    > WHERE salary>30000;
OK
Time taken: 0.592 seconds
hive>
```

```
hive> SELECT * FROM emp_30000;
```

```
hive> SELECT * FROM emp_30000;
OK
NULL    "Kranthi"       30000   "Op Admin"
Time taken: 0.185 seconds, Fetched: 1 row(s)
hive>
```

#Dropping a View

```
hive> DROP VIEW emp_30000;
```

```
hive> DROP VIEW emp_30000;
OK
Time taken: 0.153 seconds
```

Note:index has been removed from hive after 3.* update