

PRACTICAL BOOK  
MSC CS (PART II) SEMESTER - III  
2022-23

SUBJECT

Big Data Analytics  
SUBMITTED BY

**Samreen Bano Raeen**

Seat No. **CS21019**

Submitted in partial fulfilment of the requirement for  
Qualifying

M.Sc. CS Part II Semester III Examination  
2022-23

University of Mumbai  
Department of Computer Science  
R.D & S.H National & W.A Science College  
Linking Road, Bandra (w), Mumbai-50

# Big Data Analytics

# INDEX

Sr. No.	PRACTICAL	Sign
1	Installing and setting environment variables for Working with Apache Hadoop and Implementing Map-Reduce Program for Word Count problem	
2	Write a Spark code for the given application and handle error and recovery of data	
3	Install Hive and use Hive Create and store structured databases	
4	Install HBase and use the HBase Data model Store and retrieve data.	
5	Perform importing and exporting of data between SQL and Hadoop using Sqoop.	
6	Write a Pig Script for solving counting problems.	
7	Use Flume and transport the data from the various sources to a centralized data store.	



# R. D. National & W. A. Science College

Bandra (W), Mumbai – 4000 50

Department of Computer Science  
M.Sc. (CS)

## Certificate

This is to certify that **Raeen Samreen bano** of **M.Sc Part II(Sem III)** class has satisfactorily completed **7** Practical in the subject of **Big Data Analytics** as a part of M.Sc. Degree Course in Computer Science during the academic year 2022 – 2023.

Lecturer In charge

External Examiner

Head of  
Department

College Stamp

# Practical 1

**Aim:** Writing a Map-Reduce Program to Count Words.

## What is Hadoop?

Apache Hadoop is an open-source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyse massive datasets in parallel more quickly.

Hadoop consists of four main modules:

- Hadoop Distributed File System (HDFS) – A distributed file system that runs on standard or low-end hardware. HDFS provides better data throughput than traditional file systems, in addition to high fault tolerance and native support of large datasets.
- Yet Another Resource Negotiator (YARN) – Manages and monitors cluster nodes and resource usage. It schedules jobs and tasks.
- MapReduce – A framework that helps programs do the parallel computation on data. The map task takes input data and converts it into a dataset that can be computed in key value pairs. The output of the map task is consumed by reduce tasks to aggregate output and provide the desired result.
- Hadoop Common – Provides common Java libraries that can be used across all modules.

Hadoop makes it easier to use all the storage and processing capacity in cluster servers, and to execute distributed processes against huge amounts of data. Hadoop provides the building blocks on which other services and applications can be built.

Applications that collect data in various formats can place data into the Hadoop cluster by using an API operation to connect to the NameNode. The NameNode tracks the file directory structure and placement of “chunks” for each file, replicated across DataNodes. To run a job to query the data, provide a MapReduce job made up of many map and reduce tasks that run against the data in HDFS spread across the DataNodes. Map tasks run on each node against the input files supplied, and reducers run to aggregate and organize the final output.

## What is MapReduce?

MapReduce is a programming paradigm that enables massive scalability across hundreds or thousands of servers in a Hadoop cluster. As the processing component, MapReduce is the heart of Apache Hadoop. The term "MapReduce" refers to two separate and distinct tasks that Hadoop programs perform. The first is the map job, which takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs).

The reduce job takes the output from a map as input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce job is always performed after the map job.

MapReduce programming offers several benefits to help you gain valuable insights from your big data:

- Scalability: Businesses can process petabytes of data stored in the Hadoop Distributed File System (HDFS).
- Flexibility: Hadoop enables easier access to multiple sources of data and multiple types of data.
- Speed: With parallel processing and minimal data movement, Hadoop offers fast processing of massive amounts of data.
- Simple: Developers can write code in a choice of languages, including Java, C++, and Python.

## Pre-requisites:

1. Check the version of Java that is currently available on your system.

```
(kali㉿kali)-[~]
└─$ java -version
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
openjdk version "17.0.5" 2022-10-18
OpenJDK Runtime Environment (build 17.0.5+8-Debian-2)
OpenJDK 64-Bit Server VM (build 17.0.5+8-Debian-2, mixed mode, sharing)
```

2. Update the linux distribution on your system.

```
(kali㉿kali)-[~]
└─$ sudo apt-get install update
[sudo] password for kali:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
E: Unable to locate package update
```

3. Update your Java Development Kit (JDK)

```
(kali㉿kali)-[~]
└─$ sudo apt-get install default-jdk
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following packages were automatically installed and are no longer required:
libatk1.0-data libevas4 libexporter-tiny-perl libflac8 libgs9-common libgssdp-1.2-0 libgnupnp-1.2-1
libhttp-server-simple-perl libimbase25 liblerc3 liblist-moreutils-perl liblist-moreutils-xs-perl libopenexr25 libopenh264-6
libperl5.34 libplacebo192 libpoppler18 libpython3.9-minimal libpython3.9-stdlib libsvtavenc0 libwebsockets16 libwireshark15
libwiretap12 libwsutil13 openjdk-11-jre perl-modules-5.34 python3-dataclasses-json python3-limiter python3-marshmallow-enum
python3-mypy-extensions python3-responses python3-spyse python3-token-bucket python3-typing-inspect python3.9 python3.9-minimal
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
default-jdk-headless libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libxau-dev libxcb1-dev libxdmcp-dev libxt-dev
openjdk-17-jdk openjdk-17-jdk-headless x11proto-dev xorg-sgml-doctools xtrans-dev
Suggested packages:
libice-doc libsm-doc libx11-doc libxcb1-doc libxt-doc openjdk-17-demo openjdk-17-source visualvm
The following NEW packages will be installed:
default-jdk default-jdk-headless libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libxau-dev libxcb1-dev libxdmcp-dev
libxt-dev openjdk-17-jdk openjdk-17-jdk-headless x11proto-dev xorg-sgml-doctools xtrans-dev
0 upgraded, 15 newly installed, 0 to remove and 310 not upgraded.
Need to get 236 MB of archives.
After this operation, 250 MB of additional disk space will be used.
Do you want to continue? [Y/n] y
```

4. Note that, in order to use Hadoop, you require Java 8 (jdk1.8.0\_341)

Here I have downloaded and jdk1.8.0\_341 x64 Linux Distribution from Oracle.

```

(kali㉿kali)-[~]
└─$ java -version
openjdk version "17.0.5" 2022-10-18
OpenJDK Runtime Environment (build 17.0.5+8-Debian-2)
OpenJDK 64-Bit Server VM (build 17.0.5+8-Debian-2, mixed mode, sharing)

(kali㉿kali)-[~]
└─$ cd /usr/lib/jvm

(kali㉿kali)-[/usr/lib/jvm]
└─$ ls
default-java          java-1.17.0-openjdk-amd64  java-17-openjdk-amd64
java-1.11.0-openjdk-amd64  java-11-openjdk-amd64  openjdk-17

(kali㉿kali)-[/usr/lib/jvm]
└─$ sudo tar -xvzf ~/Downloads/jdk-8u341-linux-x64.tar.gz
[sudo] password for kali:
[jdk1.8.0_341/COPYRIGHT
[jdk1.8.0_341/LICENSE
[jdk1.8.0_341/README.html
[jdk1.8.0_341/THIRDPARTYLICENSEREADME.txt
[jdk1.8.0_341/bin/java_wmi.cgi

```

Also remember to add the path in the /etc/environment file where your java package is installed. Add the path “`/usr/lib/jvm/jdk1.8.0_341/bin:/usr/lib/jvm/jdk1.8.0_341/jre/bin`” after “`games:`”

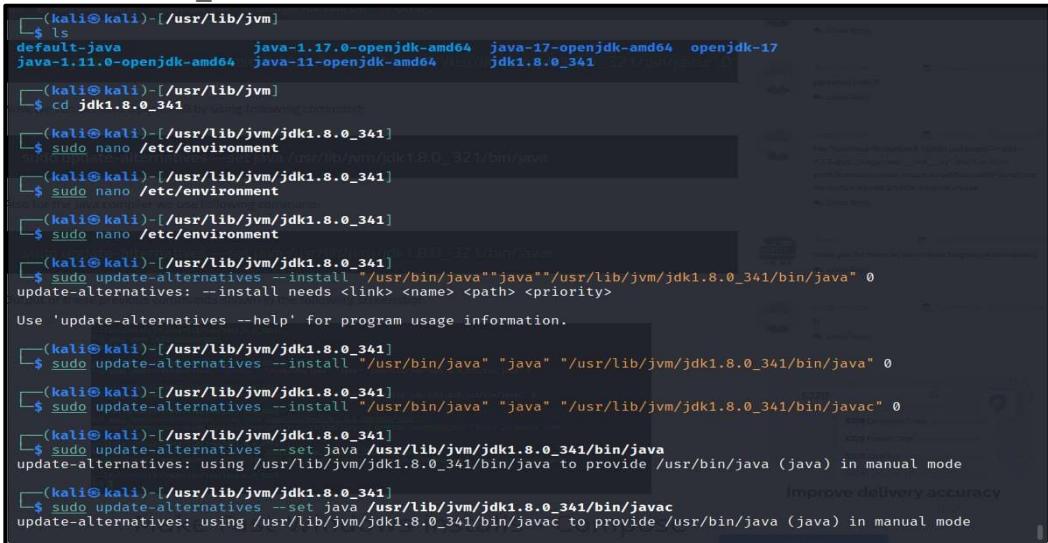


```

File Actions Edit View Help
GNU nano 6.4                               kali@kali: /etc/environment
# START KALI-DEFAULTS CONFIG
# Everything from here and until STOP_KALI-DEFAULTS CONFIG
# was installed by the kali-defaults package, and it will
# be removed if ever the kali-defaults package is removed.
# If you want to disable a line, please do NOT remove it,
# as it would be added back when kali-defaults is upgraded.
# Instead, comment the line out, and your change will be
# preserved across upgrades.
PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/bin:/usr/games:/usr/games:usr/lib/jvm/jdk1.8.0_341/bin:/usr/lib/jvm/jdk1.8.0_341/jre/bin
COMMAND_NOT_FOUND_INSTALL_PROMPT=1
POWERSHELL_UPDATECHECK=OFF
POWERSHELL_TELEMETRY_OPTOUT=1
DOTNET_CLI_TELEMETRY_OPTOUT=1
# STOP KALI-DEFAULTS CONFIG

```

I have to use the “`update-alternatives`” command to set-up the version of “`java`” and “`javac`”. Note, using the correct version of java and javac is especially important, make sure that they are both set to `1.8.0_341`.



```

(kali㉿kali)-[/usr/lib/jvm]
└─$ ls
default-java          java-1.17.0-openjdk-amd64  java-17-openjdk-amd64  openjdk-17
java-1.11.0-openjdk-amd64  java-11-openjdk-amd64  jdk1.8.0_341

(kali㉿kali)-[/usr/lib/jvm]
└─$ cd jdk1.8.0_341
by typing following command

(kali㉿kali)-[/usr/lib/jvm/jdk1.8.0_341]
└─$ sudo nano /etc/environment
by typing following command

(kali㉿kali)-[/usr/lib/jvm/jdk1.8.0_341]
└─$ sudo nano /etc/environment
by typing following command

(kali㉿kali)-[/usr/lib/jvm/jdk1.8.0_341]
└─$ sudo nano /etc/environment
by typing following command

(kali㉿kali)-[/usr/lib/jvm/jdk1.8.0_341]
└─$ sudo update-alternatives --install "/usr/bin/java" "java" "/usr/lib/jvm/jdk1.8.0_341/bin/java" 0
update-alternatives: --install needs <link> <name> <path> <priority>
Use 'update-alternatives --help' for program usage information.

(kali㉿kali)-[/usr/lib/jvm/jdk1.8.0_341]
└─$ sudo update-alternatives --install "/usr/bin/java" "java" "/usr/lib/jvm/jdk1.8.0_341/bin/java" 0
by typing following command

(kali㉿kali)-[/usr/lib/jvm/jdk1.8.0_341]
└─$ sudo update-alternatives --install "/usr/bin/javac" "javac" "/usr/lib/jvm/jdk1.8.0_341/bin/javac" 0
by typing following command

(kali㉿kali)-[/usr/lib/jvm/jdk1.8.0_341]
└─$ sudo update-alternatives --set java /usr/lib/jvm/jdk1.8.0_341/bin/java
update-alternatives: using /usr/lib/jvm/jdk1.8.0_341/bin/java to provide /usr/bin/java (java) in manual mode

(kali㉿kali)-[/usr/lib/jvm/jdk1.8.0_341]
└─$ sudo update-alternatives --set javac /usr/lib/jvm/jdk1.8.0_341/bin/javac
update-alternatives: using /usr/lib/jvm/jdk1.8.0_341/bin/javac to provide /usr/bin/javac (javac) in manual mode

```

```
(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --config java
There are 4 choices for the alternative java (providing /usr/bin/java).

Selection    Path                                Priority  Status
-----  -----
0            /usr/lib/jvm/java-17-openjdk-amd64/bin/java  1711      auto mode
1            /usr/lib/jvm/java-11-openjdk-amd64/bin/java   1111      manual mode
2            /usr/lib/jvm/java-17-openjdk-amd64/bin/java  1711      manual mode
3            /usr/lib/jvm/jdk1.8.0_341/bin/java           0        manual mode
* 4          /usr/lib/jvm/jdk1.8.0_341/bin/javac         0        manual mode

Press <enter> to keep the current choice[*], or type selection number: 3
update-alternatives: using /usr/lib/jvm/jdk1.8.0_341/bin/java to provide /usr/bin/java (java) in manual mode

(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ java -version
java version "1.8.0_341"
Java(TM) SE Runtime Environment (build 1.8.0_341-b10)
Java HotSpot(TM) 64-Bit Server VM (build 25.341-b10, mixed mode)

(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
```

```
(hadoopusr@kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --install "/usr/bin/javac" "javac" "/usr/lib/jvm/jdk1.8.0_341/bin/javac" 1

(hadoopusr@kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --set javac /usr/lib/jvm/jdk1.8.0_341/bin/javac
update-alternatives: using /usr/lib/jvm/jdk1.8.0_341/bin/javac to provide /usr/bin/javac (javac) in manual mode

(hadoopusr@kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --config javac
There are 2 choices for the alternative javac (providing /usr/bin/javac).

Selection    Path                                Priority  Status
-----  -----
0            /usr/lib/jvm/java-17-openjdk-amd64/bin/javac  1711      auto mode
1            /usr/lib/jvm/java-17-openjdk-amd64/bin/javac   1711      manual mode
* 2          /usr/lib/jvm/jdk1.8.0_341/bin/javac         1        manual mode

Press <enter> to keep the current choice[*], or type selection number: 2
```

## 5. Check if Java was successfully installed

```
(hadoopusr@kali)-[~/home/kali/Desktop/WordCountProgram]
$ java -version
javac 1.8.0_341

(hadoopusr@kali)-[~/home/kali/Desktop/WordCountProgram]
$ javac -version
javac 1.8.0_341

(hadoopusr@kali)-[~/home/kali/Desktop/WordCountProgram]
```

## 6. Create a separate user on linux to install Hadoop.

Here we will create a user “hadoopusr” in the user-group “hadoop”.

```
(kali㉿kali)-[~]
$ sudo addgroup hadoop
Adding group `hadoop' (GID 1001) ...
Done.

(kali㉿kali)-[~]
$ sudo adduser --ingroup hadoop hadoopusr
Adding user `hadoopusr' ...
Adding new user `hadoopusr' (1002) with group `hadoop (1001)' ...
Creating home directory `/home/hadoopusr' ...
Copying files from `/etc/skel' ...
New password:
Retype new password:
passwd: password updated successfully
Changing the user information for hadoopusr
Enter the new value, or press ENTER for the default
    Full Name []: Big Data Project
    Room Number []:
    Work Phone []:
    Home Phone []:
    Other []:
Is the information correct? [Y/n] Y
Adding new user `hadoopusr' to supplemental / extra groups `users' ...
Adding user `hadoopusr' to group `users' ...

(kali㉿kali)-[~]
$
```

```
(kali㉿kali)-[~]
$ sudo adduser hadoopusr sudo
Adding user `hadoopusr' to group `sudo' ...
Done.
```

## 7. Now we will install the SSH Server.

This is necessary in order for us to start the various components in Hadoop.

```
(kali㉿kali)-[~]
$ sudo apt-get install openssh-server
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
openssh-server is already the newest version (1:9.0p1-1+b2).
openssh-server set to manually installed.
The following packages were automatically installed and are no longer required:
libatck1.0-data libev4 libexporter-tiny-perl libflac8 libgs9-common libgssdp-1.2-0 libgupnp-1.2-1
libhttp-server-simple-perl liblilmbase25 liblerc3 liblist-moreutils-perl liblist-moreutils-xs-perl libopenh264-6
libperl5.34 libplacebo192 libpoppler118 libpython3.9-minimal libpython3.9-stdlib libsavvienic0 libwebsocket16 libwireshark15
libwiretap12 libwsutil13 openjdk-11-jre perl-modules-5.34 python3-dataclasses-json python3-limiter python3-marshmallow-enum
python3-mypy-extensions python3-responses python3-spyce python3-token-bucket python3-typing-inspect python3.9 python3.9-minimal
Use 'sudo apt autoremove' to remove them.
0 upgraded, 0 newly installed, 0 to remove and 310 not upgraded.
```

## 8. Now we will switch users from “kali” to “hadoopusr”

```
(kali㉿kali)-[~]
$ su - hadoopusr
Password:
(hadoopusr㉿kali)-[~]
$
```

9. Then we will create a SSH key.

```
(hadoopusr@kali)-[~]
$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoopusr/.ssh/id_rsa):
Created directory '/home/hadoopusr/.ssh'.
Your identification has been saved in /home/hadoopusr/.ssh/id_rsa.
Your public key has been saved in /home/hadoopusr/.ssh/id_rsa.pub.
The key fingerprint is:
SHA256:IWk+5eEh0066obQYPf52gTu+tATEXeH7rqx9DHeA0EI hadoopusr@kali
The key's randomart image is:
+---[RSA 3072]---+
| .E.o. |
| . .oo.. |
| o .o*.= |
| . 0.0.+ |
| . .B S. |
| . ..+0+. . |
| . + =..= .. |
| = *+*..0 |
| . ++B=+o. |
+---[SHA256]---+
(hadoopusr@kali)-[~]
$
```

10. Then we have to save the key.

```
(hadoopusr@kali)-[~]
$ cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
-bash: /home/hadoopusr/: Is a directory
```

11. Now, we will start a SSH server on localhost.

```
(hadoopusr@kali)-[~]
$ sudo service ssh status
[sudo] password for hadoopusr:
● ssh.service - OpenBSD Secure Shell server
  Loaded: loaded (/lib/systemd/system/ssh.service; disabled; preset: disabled)
  Active: inactive (dead)
    Docs: man:sshd(8)
          man:sshd_config(5)

(hadoopusr@kali)-[~]
$ sudo service ssh start

(hadoopusr@kali)-[~]
$ sudo service ssh status
● ssh.service - OpenBSD Secure Shell server
  Loaded: loaded (/lib/systemd/system/ssh.service; disabled; preset: disabled)
  Active: active (running) since Sun 2022-11-27 02:19:55 EST; 2s ago
    Docs: man:sshd(8)
          man:sshd_config(5)
  Process: 9820 ExecStartPre=/usr/sbin/sshd -t (code=exited, status=0/SUCCESS)
  Main PID: 9821 (sshd)
    Tasks: 1 (limit: 2281)
   Memory: 2.8M
      CPU: 27ms
     CGroup: /system.slice/ssh.service
             └─9821 "sshd: /usr/sbin/sshd -D [listener] 0 of 10-100 startups"

Nov 27 02:19:55 kali systemd[1]: Starting OpenBSD Secure Shell server...
Nov 27 02:19:55 kali sshd[9821]: Server listening on 0.0.0.0 port 22.
Nov 27 02:19:55 kali sshd[9821]: Server listening on :: port 22.
Nov 27 02:19:55 kali systemd[1]: Started OpenBSD Secure Shell server.
```

```
(hadoopusr㉿kali)-[~]
$ ssh localhost
The authenticity of host 'localhost (::1)' can't be established.
ED25519 key fingerprint is SHA256:d90PVKUb7nPhTRKzzQStWW/1r3757yecGCwlhyXJys4.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ED25519) to the list of known hosts.
hadoopusr@localhost's password:
Linux kali 5.19.0-kali2-amd64 #1 SMP PREEMPT_DYNAMIC Debian 5.19.11-1kali2 (2022-10-10) x86_64

The programs included with the Kali GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*copyright.

Kali GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.
(hadoopusr㉿kali)-[~]
$
```

12. Now we will exit the “hadoopusr”

```
(hadoopusr㉿kali)-[~]
$ exit
logout
Connection to localhost closed.

(hadoopusr㉿kali)-[~]
$
```

13. Next. We will download Apache Hadoop 2.9.0.

The screenshot shows a web browser window with the URL <https://archive.apache.org/dist/hadoop/common/hadoop-2.9.0/>. The page title is "Index of /dist/hadoop/common/hadoop-2.9.0". Below the title is a table listing various files and their details:

Name	Last modified	Size	Description
Parent Directory		-	
<a href="#">hadoop-2.9.0-src.tar.gz</a>	2017-11-17 23:08	37M	
<a href="#">hadoop-2.9.0-src.tar.gz.asc</a>	2017-11-17 23:08	819	
<a href="#">hadoop-2.9.0-src.tar.gz.md5</a>	2017-11-17 23:08	162	
<a href="#">hadoop-2.9.0-src.tar.gz.mds</a>	2017-11-17 23:08	1.0K	
<a href="#">hadoop-2.9.0-src.tar.gz.sha256</a>	2018-03-13 20:32	90	
<a href="#">hadoop-2.9.0.tar.gz</a>	2017-11-17 23:10	350M	
<a href="#">hadoop-2.9.0.tar.gz.asc</a>	2017-11-17 23:08	819	
<a href="#">hadoop-2.9.0.tar.gz.md5</a>	2017-11-17 23:08	154	
<a href="#">hadoop-2.9.0.tar.gz.mds</a>	2017-11-17 23:08	1.0K	
<a href="#">hadoop-2.9.0.tar.gz.sha256</a>	2018-03-13 20:32	86	

14. Then we will install it.

```
(hadoopusr㉿kali)-[~]
$ cd /home/kali/Desktop
(hadoopusr㉿kali)-[/home/kali/Desktop]
$ ls
hadoop-2.9.0.tar.gz
```

```
(hadoopusr㉿kali)-[~/home/kali/Desktop]
└─$ sudo tar -xvzf hadoop-2.9.0.tar.gz
hadoop-2.9.0/
hadoop-2.9.0/include/
hadoop-2.9.0/include/Pipes.hh
hadoop-2.9.0/include/SerialUtils.hh
hadoop-2.9.0/include/hdfs.h
hadoop-2.9.0/include/StringUtils.hh
hadoop-2.9.0/include/TemplateFactory.hh
hadoop-2.9.0/NOTICE.txt
hadoop-2.9.0/lib/
hadoop-2.9.0/lib/native/
hadoop-2.9.0/lib/native/libhdfs.a
hadoop-2.9.0/lib/native/libhadoop.so
hadoop-2.9.0/lib/native/libhadoop.so.1.0.0
hadoop-2.9.0/lib/native/libhdfs.so.0.0.0
hadoop-2.9.0/lib/native/libhadoop.a
hadoop-2.9.0/lib/native/libhdfs.so
hadoop-2.9.0/lib/native/examples/
```

15. Now we will move the file to the “hadoopusr” and transfer ownership to the “hadoopusr”.

```
(hadoopusr㉿kali)-[~/home/kali/Desktop]
└─$ ls
hadoop-2.9.0  hadoop-2.9.0.tar.gz

(hadoopusr㉿kali)-[~/home/kali/Desktop]
└─$ sudo mv hadoop-2.9.0 /usr/local/hadoop

(hadoopusr㉿kali)-[~/home/kali/Desktop]
└─$ sudo chown -R hadoopusr /usr/local/hadoop

(hadoopusr㉿kali)-[~/home/kali/Desktop]
└─$ █
```

16. Make the configurations to the “bashrc” file using nano

```
(hadoopusr㉿kali)-[~/home/kali/Desktop]
└─$ sudo nano ~/.bashrc
```

The Configurations are:

```
export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_341 export JAVA_PATH=$PATH:$JAVA_HOME/bin
export HADOOP_HOME=/usr/local/hadoop export PATH=$PATH:$HADOOP_HOME/bin export
PATH=$PATH:$HADOOP_HOME/sbin 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=$HADOOP_HOME/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/native"
```

```

# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

if [ -f ~/.bash_aliases ]; then
  . ~/.bash_aliases
fi

# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile).
# sources /etc/bash.bashrc.

if ! shopt -oq posix; then
  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

export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64
export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$HADOOP_HOME/bin
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=$HADOOP_HOME/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/native"

^G Help      ^O Write Out   ^W Where Is   ^K Cut        ^T Execute   ^C Location   M-U Undo   M-A Set Mark
^X Exit      ^R Read File   ^N Replace    ^U Paste     ^J Justify   ^Y Go To Line  M-E Redo   M-C Copy

```

Then we will run “source ~/.bashrc” to enable the changes.

```

(hadoopusr@kali)-[~/home/kali/Desktop]
$ source ~/.bashrc
(hadoopusr@kali)-[~/home/kali/Desktop]
$ 

```

17. Next we will make some changes to the hadoop environment file

We will run the command “sudo nano /usr/local/hadoop/etc/hadoop/hadoop-env.sh”

```

(hadoopusr@kali)-[~/home/kali/Desktop]
$ sudo nano /usr/local/hadoop/etc/hadoop/hadoop-env.sh

```

Here we will set the path of JAVA\_HOME to “JAVA\_HOME=/usr/lib/jvm/jdk1.8.0\_341”.

```

# The java implementation to use.
#export JAVA_HOME=${JAVA_HOME}
export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_341

# You will get the following response:
# The jsvc implementation to use. Jsvc is required to run secure datanodes
# that bind to privileged ports to provide authentication of data transfer
# protocol. Jsvc is not required if SASL is configured for authentication of
# data transfer protocol using non-privileged ports.
#export JSVC_HOME=${JSVC_HOME}

export HADOOP_CONF_DIR=${HADOOP_CONF_DIR:-"/etc/hadoop"}

# Extra Java CLASSPATH elements. Automatically insert capacity-scheduler.
for f in $HADOOP_HOME/contrib/capacity-scheduler/*.jar; do
  if [ "$HADOOP_CLASSPATH" ]; then
    export HADOOP_CLASSPATH=$HADOOP_CLASSPATH:$f
  else
    export HADOOP_CLASSPATH=$f
  fi
done >>> and extract its contents using the following commands.
done

# The maximum amount of heap to use, in MB. Default is 1000.
#export HADOOP_HEAPSIZE=
#export HADOOP_NAMENODE_INIT_HEAPSIZE=""

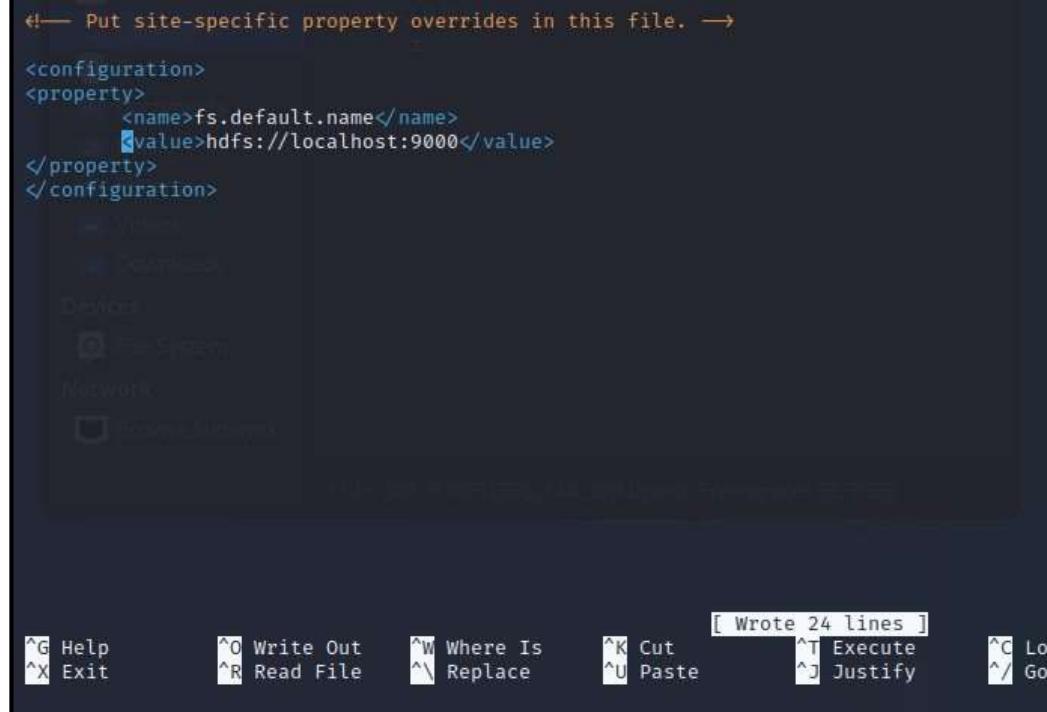
^G Help      ^O Write Out   ^W Where Is   ^K Cut        ^T Execute   ^C Location   M-U Undo   M-A Set Mark
^X Exit      ^R Read File   ^N Replace    ^U Paste     ^J Justify   ^Y Go To Line  M-E Redo   M-C Copy

```

18. Now we will make some changes to the core-site.xml file.

Core-Site.xml configuration:

```
<property>
  <name>fs.default.name</name>
  <value>hdfs://localhost:9000</value>
</property>
```



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

<configuration>
<property>
  <name>fs.default.name</name>
  <value>hdfs://localhost:9000</value>
</property>
</configuration>
```

19. Now we will make some changes to the hdfs-site.xml HDFS-Site.xml configuration:

```
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
<property>
  <name>dfs.name.name.dir</name>
  <value>file:/usr/local/hadoop_tmp/hdfs/namenode</value>
</property>
<property>
  <name>dfs.data.data.dir</name>
<value>file:/usr/local/hadoop_tmp/hdfs/datanode</value> </property>
```

```

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

<configuration>
<property>
    <name>dfs.replication</name>
    <value>1</value>
</property>
<property>
    <name>dfs.name.name.dir</name>
    <value>file:/usr/local/hadoop_tmp/hdfs/namenode</value>
</property>
<property>
    <name>dfs.data.data.dir</name>
    <value>file:/usr/local/hadoop_tmp/hdfs/datanode</value>
</property>
</configuration>

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

```

20. Now we will configure the yarn-site.xml YARN-Site.xml configuration:

```

<property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>

```

```
</property>
```

```

<configuration>
<property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
</property>
<!-- Site specific YARN configuration properties -->
</configuration>

```

21. Now we will configure the mapred-site.xml Mapred-site.xml configuration:

```

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

```

```
</property>
```

```

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

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

```

22. Next, we will create the following directories and transfer the ownership to the “hadoopusr”.

These directories are where hadoop will store data about the namenode and the datanode.

```
(hadoopusr@kali)-[~/Desktop]
$ sudo mkdir -p /usr/local/hadoop_space

(hadoopusr@kali)-[~/Desktop]
$ sudo mkdir -p /usr/local/hadoop_space/hdfs/namenode

(hadoopusr@kali)-[~/Desktop]
$ sudo mkdir -p /usr/local/hadoop_space/hdfs/datanode

(hadoopusr@kali)-[~/Desktop]
$ sudo chown -R hadoopusr /usr/local/hadoop_space

(hadoopusr@kali)-[~/Desktop]
```

23. Then we format the namenode in the HDFS file structure.

```
(hadoopusr@kali)-[~]
$ hdfs namenode -format
22/11/27 03:34:09 INFO namenode.NameNode: STARTUP_MSG:
/*****STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = kali/127.0.1.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 2.9.0
STARTUP_MSG: classpath = /usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/hadoop/common/lib/java-xml
builder-0.4.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-digester-1.8.jar:/usr/local/hadoop/share
/hadoop/common/lib/slf4j-api-1.7.25.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-net-3.1.jar:/usr
/local/hadoop/share/hadoop/common/lib/mockito-all-1.8.5.jar:/usr/local/hadoop/share/hadoop/common/lib/gson
-2.2.4.jar:/usr/local/hadoop/share/hadoop/common/lib/snappy-java-1.0.5.jar:/usr/local/hadoop/share/hadoop/
common/lib/commons-codec-1.4.jar:/usr/local/hadoop/share/hadoop/common/lib/xz-1.0.jar:/usr/local/hadoop/sh
are/hadoop/common/lib/jettison-1.1.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-lang-2.6.jar:/usr
```

24. Next we will start the Distributed file system.

```
(hadoopusr@kali)-[~]
$ start-dfs.sh
22/11/27 03:36:31 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... us
ing builtin-java classes where applicable
Starting namenodes on [localhost]
hadoopusr@localhost's password:
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hadoopusr-namenode-kali.out
hadoopusr@localhost's password:
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hadoopusr-datanode-kali.out
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
ED25519 key fingerprint is SHA256:d90PVKUb7nPhTRKzzQ5tWW/1r3757yecGCwlhyXJys4.
This host key is known by the following other names/addresses:
  ~/.ssh/known_hosts:1: [hashed name]
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
0.0.0.0: Warning: Permanently added '0.0.0.0' (ED25519) to the list of known hosts.
hadoopusr@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hadoopusr-secondarynamenode-
kali.out
22/11/27 03:37:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... us
ing builtin-java classes where applicable
```

25. Followed by the YARN system.

```
(hadoopusr@kali)-[~]
$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hadoopusr-resourcemanager-kali.out
hadoopusr@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hadoopusr-nodemanager-kali.out
```

26. Check if all the systems are running

```
(hadoopusr@kali)-[~]
└─$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
22/11/27 03:43:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
hadoopusr@localhost's password:
localhost: namenode running as process 29427. Stop it first.
hadoopusr@localhost's password:
localhost: datanode running as process 29636. Stop it first.
Starting secondary namenodes [0.0.0.0]
hadoopusr@0.0.0.0's password:
0.0.0.0: secondarynamenode running as process 29875. Stop it first.
22/11/27 03:43:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hadoopusr-resourcemanager-kali.out
hadoopusr@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hadoopusr-nodemanager-kali.out

(hadoopusr@kali)-[~]
└─$ jps
29875 SecondaryNameNode
29427 NameNode
29636 DataNode
32859 Jps
```

27. Here we can see that hadoop has been successfully installed and activated We have to go the url <http://localhost:500070> to access the Hadoop Server page.

**Overview 'localhost:9000' (active)**

Started:	Sun Nov 27 03:36:58 -0500 2022
Version:	2.9.0, r756ebc8394e473ac25feac05fa493f6d612e6c50
Compiled:	Mon Nov 13 18:15:00 -0500 2017 by arsuresh from branch-2.9.0
Cluster ID:	CID-c1d1539a-8faa-4206-bdd2-8ab3a5efe94d
Block Pool ID:	BP-1309047964-127.0.1.1-1669538050401

**Summary**

Security is off.  
Safemode is off.  
1 files and directories, 0 blocks = 1 total filesystem object(s).  
Heap Memory used 39.86 MB of 64 MB Heap Memory. Max Heap Memory is 1000 MB.  
Non Heap Memory used 43.62 MB of 46.19 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

Configured Capacity:	78.28 GB
DFS Used:	24 KB (0%)
Non DFS Used:	18.95 GB
DFS Remaining:	55.31 GB (70.65%)

We can also see the information of the cluster.

**All Applications**

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus	Running Containers	Allocated CPU vCores	Allocated Memory MB	Reserved CPU vCores	Reserved Memory MB	% of Queue	% of Cluster	Progress	Tracking UI	Blacklisted Nodes
No data available in table																			

28. Now we will perform the Word-Count program on Hadoop MapReduce. First we will check the hadoop version.

```
(hadoopusr@kali)-[~]
$ hadoop version
Hadoop 2.9.0
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r 756ebc8394e473ac25feac05fa493f6d612e6c50
Compiled by arsuresh on 2017-11-13T23:15Z
Compiled with protoc 2.5.0
From source with checksum 0a76a9a32a5257331741f8d5932f183
This command was run using /usr/local/hadoop/share/hadoop/common/hadoop-common-2.9.0.jar
```

29. Then we make sure that java and javac is running properly.



30. Next we will write the WordCount program in java.

Code:

```
import java.io.IOException;
import java.util.StringTokenizer;
```

```
import org.apache.hadoop.conf.Configuration; import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable; import org.apache.hadoop.io.Text; import
org.apache.hadoop.mapreduce.Job; import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer; import
org.apache.hadoop.mapreduce.lib.input.FileInputFormat; import
org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
```

```
public class WordCount {

    public static class TokenizerMapper
        extends Mapper<Object, Text, Text, IntWritable>{

        private final static IntWritable one = new IntWritable(1);    private Text word = new
Text();

        public void map(Object key, Text value, Context context ) throws
IOException, InterruptedException {    StringTokenizer itr = new
StringTokenizer(value.toString());    while (itr.hasMoreTokens()) {
word.set(itr.nextToken());    context.write(word, one);

    }
}

    }

    public static class IntSumReducer    extends
Reducer<Text,IntWritable,Text,IntWritable> {
        private IntWritable result = new IntWritable();
        public void reduce(Text key, Iterable<IntWritable> values,
```

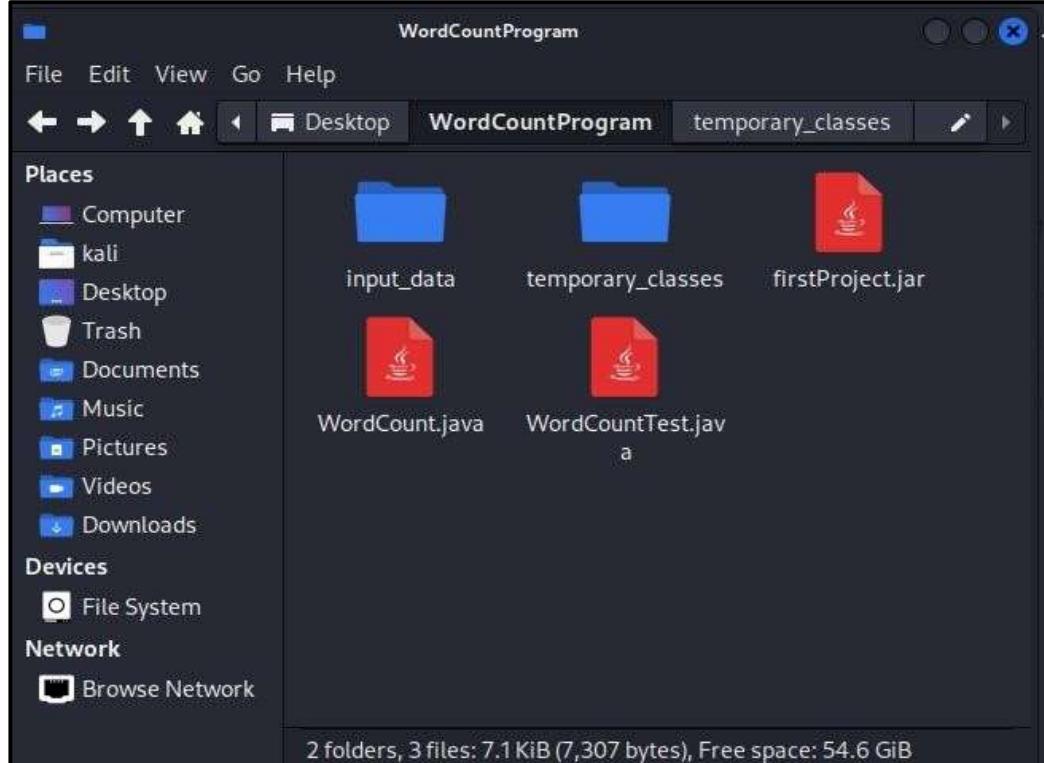
```

        Context context
    ) throws IOException, InterruptedException {    int sum = 0;
for (IntWritable val : values) {      sum += val.get();
    }    result.set(sum);
    context.write(key, result);
}
}

public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();    Job job =
Job.getInstance(conf, "word count");    job.setJarByClass(WordCount.class);
job.setMapperClass(TokenizerMapper.class);
job.setCombinerClass(IntSumReducer.class);
job.setReducerClass(IntSumReducer.class);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
FileInputFormat.addInputPath(job, new Path(args[0]));
FileOutputFormat.setOutputPath(job, new Path(args[1]));
System.exit(job.waitForCompletion(true) ? 0 : 1);
}
}
}

```

31. Then we create a folder input\_data and temporary\_classes.



32. Then we create a text file in the input\_data folder which will contain our input values The Input: Cars, Sleep, School, Apple, Dog, Speed, Blast, Cars, Mumbai, Delhi, Apple, Sleep, Car, Car, Sleep, Sleep, School.

33. Then we will set the class path for Hadoop.

```
(hadoopusr@kali)-[~]
$ export HADOOP_CLASSPATH=$(hadoop classpath)

(hadoopusr@kali)-[~]
$ echo $HADOOP_CLASSPATH
/usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/hadoop/common/lib/*:/usr/local/hadoop/share/hadoop/common/*:/usr/local/hadoop/share/hadoop/dfs:/usr/local/hadoop/share/hadoop/dfs/lib/*:/usr/local/hadoop/share/hadoop/dfs/*:/usr/local/hadoop/share/hadoop/yarn:/usr/local/hadoop/share/hadoop/yarn/lib/*:/usr/local/hadoop/share/hadoop/yarn/*:/usr/local/hadoop/share/hadoop/mapreduce/lib/*:/usr/local/hadoop/share/hadoop/mapreduce/*:/usr/local/hadoop/contrib/capacity-scheduler/*.jar
```

34. Then we will create a directory on the HDFS.

```
(hadoopusr@kali)-[~]
$ hadoop fs -mkdir /WordCountExample
22/11/27 04:07:28 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

35. Then we will create a directory inside the folder WordCountExample for the input file.

```
(hadoopusr@kali)-[~]
$ hadoop fs -mkdir /WordCountExample/Input
22/11/27 04:09:02 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

(hadoopusr@kali)-[~]
```

36. Then we check the filesystem on Hadoop Administration page that is accessible at localhost:50070

The screenshot shows the Hadoop Administration interface with the 'Overview' tab selected. The URL is 'localhost:50070/dfshealth.html#tab-overview'. The page displays basic cluster statistics and links to 'Browse the file system' and 'Logs'.

The screenshot shows the 'Browse Directory' page with the path '/'. It lists a single entry: 'WordCountProject' (drwxr-xr-x, hadoopusr, supergroup, 0 B, Nov 28 04:34). The page includes search and filter options, and navigation buttons for 'Previous', '1', and 'Next'.

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxr-xr-x	hadoopusr	supergroup	0 B	Nov 28 04:34	0	0 B	WordCountProject

## Browse Directory

/WordCountProject									Go!	File	Upload	Folder
Show 25 entries		Search:										
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	Actions				
drwxr-xr-x	hadoopusr	supergroup	0 B	Nov 28 04:35	0	0 B	Input					

Hadoop, 2017.

## 37. Now we will add our input file to the hadoop server

```
(hadoopusr@kali)-[~]
$ hadoop fs -put '/home/kali/Desktop/WordCountProgram/input_data/input.txt' /WordCountProject/Input
22/11/28 04:39:43 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
(hadoopusr@kali)-[~]
$
```

## Hadoop

Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities

### Browse Directory

/WordCountProject/Input									Go!	File	Upload	Folder
Show 25 entries		Search:										
Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name	Actions				
-rw-r--r--	hadoopusr	supergroup	97 B	Nov 28 04:39	1	128 MB	input.txt					

Hadoop, 2017.

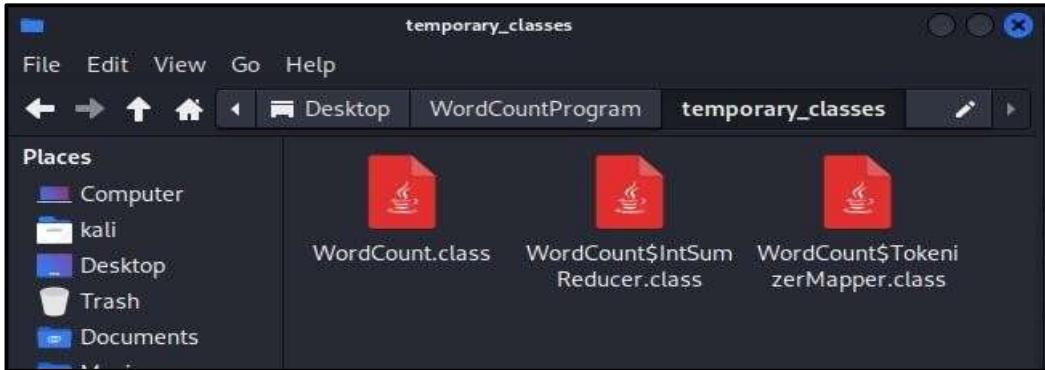
## 38. Next, we will change our directory to the Project Folder.

```
(hadoopusr@kali)-[~]
$ cd /home/kali/Desktop/WordCountProgram

(hadoopusr@kali)-[/home/kali/Desktop/WordCountProgram]
$
```

## 39. Now we will compile the java code and create a jar file

```
(hadoopusr@kali)-[~/home/kali/Desktop/WordCountProgram]
$ sudo javac -classpath ${HADOOP_CLASSPATH} -d ~/home/kali/Desktop/WordCountProgram/temporary_classes' '/home/kali/Desktop/WordCountProgram/WordCount.java'
(hadoopusr@kali)-[~/home/kali/Desktop/WordCountProgram]
$ sudo jar -cvf firstProject.jar -C temporary_classes/
added manifest
adding: WordCount$IntSumReducer.class(in = 1755) (out= 751)(deflated 57%)
adding: WordCount$TokenizerMapper.class(in = 1752) (out= 769)(deflated 56%)
adding: WordCount.class(in = 1511) (out= 833)(deflated 44%)
(hadoopusr@kali)-[~/home/kali/Desktop/WordCountProgram]
```



#### 40. Next, we will use hadoop to perform a MapReduce operation

Here we will use the jar file to execute the WordCount program on Hadoop.

```
(hadoopusr@kali)-[~/home/kali/Desktop/WordCountProgram]
$ hadoop jar '/home/kali/Desktop/WordCountProgram/firstProject.jar' WordCount /WordCountProject/Input /WordCountProject/Output
22/11/28 05:59:29 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
22/11/28 05:59:30 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/11/28 05:59:31 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface
and execute your application with ToolRunner to remedy this.
22/11/28 05:59:31 INFO input.FileInputFormat: Total input files to process : 1
22/11/28 05:59:31 INFO mapreduce.JobSubmitter: number of splits:1
22/11/28 05:59:31 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use
yarn.system-metrics-publisher.enabled
22/11/28 05:59:31 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1669627693237_0001
22/11/28 05:59:32 INFO impl.YarnClientImpl: Submitted application application_1669627693237_0001
22/11/28 05:59:32 INFO mapreduce.Job: The url to track the job: http://kali:8088/proxy/application_1669627693237_0001/
22/11/28 05:59:41 INFO mapreduce.Job: Running job: job_1669627693237_0001
22/11/28 05:59:41 INFO mapreduce.Job: Job job_1669627693237_0001 running in uber mode : false
22/11/28 05:59:46 INFO mapreduce.Job: map 0% reduce 0%
22/11/28 05:59:53 INFO mapreduce.Job: map 100% reduce 100%
22/11/28 05:59:53 INFO mapreduce.Job: Job job_1669627693237_0001 completed successfully
22/11/28 05:59:53 INFO mapreduce.Job: Counters: 49
    File System Counters
        FILE: Number of bytes read=123
        FILE: Number of bytes written=404007
        FILE: Number of read operations=0
```

```
Spilled Records=20
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=117
CPU time spent (ms)=1160
Physical memory (bytes) snapshot=494956544
Virtual memory (bytes) snapshot=3898376192
Total committed heap usage (bytes)=299892736
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=97
File Output Format Counters
Bytes Written=77

(hadoopusr@kali)-[~/home/kali/Desktop/WordCountProgram]
```

#### 41. Now we will check the output.

```
(hadoopuser㉿kali)-[~/home/kali/Desktop/WordCountProgram]
$ hadoop dfs -cat /WordCountProject/Output/
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

22/11/28 06:03:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes w
here applicable
Apple 2
Blast 1
Car 2
Cars 2
Delhi 1
Dog 1
Mumbai 1
School 2
Sleep 4
Speed 1

(hadoopuser㉿kali)-[~/home/kali/Desktop/WordCountProgram]
$
```

As we have we can see in the above output, MapReduce has been performed on the input text file, here we can see the output of the WordCount program that just shows the number of times a word has occurred in a given input set.

42. We can also see the final output in the Hadoop Administrator page.

	Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
<input type="checkbox"/>	-rw-r--r--	hadoopusr	supergroup	0 B	Nov 28 05:59	1	128 MB	_SUCCESS
<input type="checkbox"/>	-rw-r--r--	hadoopusr	supergroup	77 B	Nov 28 05:59	1	128 MB	part-r-00000

## Practical 2

**Aim:** Write a Spark code for the given application and handle error and recovery of data. Write a Spark code to Handle the Streaming of data.

**Apache Spark:** is a powerful tool for real-time analysis and ML model building, which was a research output of a class project at UC Berkeley in 2010. Spark is not a programming language like Python or Java, it is a distributed data processing framework — useful for big data processing for scalability and computational power.

**PySpark:** is an interface for Apache Spark that allows users to write Spark applications using python APIs as well as to interactively analyze big data in distributed systems. Most of the Spark features are available in PySpark such as Spark SQL, DataFrame, Streaming, MLlib (Machine Learning Library), and Spark Core.

**Google Colaboratory:** is a free Jupyter Notebook environment with many pre-installed libraries like **PySpark**, Tensorflow, Pytorch, Keras, OpenCV, and many more. It is one of the cloud services that support GPU and TPU for free. Importing a dataset and training models on the data in the Colab facilitate coding experience. We can apply different ways to import and download data in Colab.

- **Installing PySpark on Colab**

Previously, setting up PySpark required installing Java, Spark, and Hadoop into the system. However, currently, an easier and simple way is available • **pip install PySpark**.

```
+ Code + Text
[1]: pip install PySpark
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting PySpark
  Downloading pyspark-3.3.1.tar.gz (281.4 MB)
    |████████| 281.4 MB 46 kB/s
Collecting py4j==0.10.9.5
  Downloading py4j-0.10.9.5-py2.py3-none-any.whl (199 kB)
    |████████| 199 kB 52.2 MB/s
Building wheels for collected packages: PySpark
  Building wheel for PySpark (setup.py) ... done
    Created wheel for PySpark: filename=pyspark-3.3.1-py2.py3-none-any.whl size=281845512 sha256=f79c563d9b447340e01c88dd74e0e87358b9ef2a4602d75531718efd2c3b3754
    Stored in directory: /root/.cache/pip/wheels/42/59/f5/79a5bf931714dc201b26025347785f087370a10a3329a899c
Successfully built PySpark
Installing collected packages: py4j, PySpark
Successfully installed PySpark-3.3.1 py4j-0.10.9.5
```

- **Spark Session**

is the entry point for the programming Spark with dataset and DataFrame API.

```
from pyspark.sql import SparkSession
spark = SparkSession.builder\
```

```
.master("local")\  
.appName("Colab")\  
.config('spark.ui.port', '4050')\  
.getOrCreate()
```

+ Code + Text

[3] from pyspark.sql import SparkSession  
spark = SparkSession.builder\  
.master("local")\  
.appName("Colab")\  
.config('spark.ui.port', '4050')\  
.getOrCreate()

✓ 7s [4] spark

SparkSession - in-memory  
SparkContext  
[Spark UI](#)  
Version v3.3.1  
Master local  
AppName Colab

- **Loading data into PySpark**

I had saved the .csv in Google drive and loaded it from there

[5] df=spark.read.csv("/content/drive/MyDrive/PySpark/ramen\_rating.csv",header=True)

The screenshot shows the Google Colab interface. On the left, there's a sidebar with a 'Files' section showing a directory structure under 'MyDrive'. The main area has a code editor with two cells. Cell [4] contains code to initialize a SparkSession and print its configuration. Cell [5] reads a CSV file named 'ramen\_rating.csv' and shows its first 5 rows. The output of cell [5] is a table:

	Brand	Style	Country	Stars
1	New Touch	Cup	Japan	3.75
2	Just Way	Pack	Taiwan	1
3	Nissin	Cup	USA	2.25
4	Wei Lih	Pack	Taiwan	2.75
5	Ching's Secret	Pack	India	3.75

Text below the table says 'only showing top 5 rows'.

At the bottom, it shows '0s completed at 12:31 PM'.

We can print out the schema in tree format, which shows the datatypes as well:

The screenshot shows the Google Colab interface. A code cell contains the command `df.printSchema()`. The output shows the schema of the DataFrame 'df' in a tree format:

```
root
|-- Brand: string (nullable = true)
|-- Style: string (nullable = true)
|-- Country: string (nullable = true)
|-- Stars: string (nullable = true)
```

- OPERATIONS

Once the Data is loaded and DataFrame is ready we can perform various operations on it.

**1. Filter**

```
df.filter('Stars < 3').show()
```

	Brand Style	Country Stars	
1	Just Way  Pack	Taiwan  1	
	Nissin  Cup	USA  2.25	
	Wei Lih  Pack	Taiwan  2.75	
	Ripe'n'Dry  Pack	Japan  0.25	
	KOKA  Pack	Singapore  2.5	
	Lipton  Box	USA  1.5	
	Acecook  Tray	Japan  1.5	
	Ottogi  Pack	South Korea  2	
	Uni-President  Pack	Vietnam  0	
	Guava Story  Tray	South Korea  1	
	Sichuan Guangyou  Bowl	China  0	
	Indomie  Pack	Nigeria  1.5	
	Tokyo Noodle  Pack	Japan  2	
	Wang  Bowl	South Korea  2	
	Ajinotori  Pack	Japan  2.5	
	Dr. McDougall's  Cup	USA  0	
	Nissin  Cup	Hong Kong  2.5	
	Sichuan Guangyou  Pack	China  0.25	
	Liang Cheng Mai  Tray	China  1	
	Nongshim  Cup	South Korea  0.5	
+-----+-----+-----+-----+			
only showing top 20 rows			

**2. Group By**

```
df.groupby('style').count().show()
```

style count
Bowl  481
Bar  1
Box  6
null  2
Pack  1531
Cup  450
Tray  108
Can  1

**3. Rename Column Names**

```
✓ 0s df = df.withColumnRenamed('Stars','Ratings')
df.show()

+-----+-----+-----+
|   Brand|Style|   Country|Ratings|
+-----+-----+-----+
| New Touch| Cup|   Japan| 3.75|
| Just Way| Pack| Taiwan|    1|
| Nissin| Cup|   USA| 2.25|
| Wei Lih| Pack| Taiwan| 2.75|
| Ching's Secret| Pack| India| 3.75|
| Samyang Foods| Pack| South Korea| 4.75|
| Acecook| Cup| Japan|    4|
| Ikeda Shoku| Tray| Japan| 3.75|
| Ripe'n'Dry| Pack| Japan| 0.25|
| KOKA| Pack| Singapore| 2.5|
| Tao Kae Noi| Pack| Thailand|    5|
| Yamachan| Pack| USA|    5|
| Nongshim| Pack| South Korea| 4.25|
| Nissin| Bowl| Japan| 4.5|
| Nissin| Pack| Hong Kong|    5|
| KOKA| Cup| Singapore| 3.5|
| TRDP| Pack| India| 3.75|
| Yamachan| Pack| USA|    5|
| Binh Tay| Pack| Vietnam|    4|
| Paldo| Pack| South Korea| 4|
+-----+-----+-----+
only showing top 20 rows
```

#### 4. calculating the % average based on Ratings column

```
[28] df = df.withColumn('Percentage',df.Ratings*20)
df.show()

+-----+-----+-----+-----+
|   Brand|Style|   Country|Ratings|Percentage|
+-----+-----+-----+-----+
| New Touch| Cup|   Japan| 3.75|    75.0|
| Just Way| Pack| Taiwan|    1|    20.0|
| Nissin| Cup|   USA| 2.25|    45.0|
| Wei Lih| Pack| Taiwan| 2.75|    55.0|
| Ching's Secret| Pack| India| 3.75|    75.0|
| Samyang Foods| Pack| South Korea| 4.75|    95.0|
| Acecook| Cup| Japan|    4|    80.0|
| Ikeda Shoku| Tray| Japan| 3.75|    75.0|
| Ripe'n'Dry| Pack| Japan| 0.25|     5.0|
| KOKA| Pack| Singapore| 2.5|    50.0|
| Tao Kae Noi| Pack| Thailand|    5|   100.0|
| Yamachan| Pack| USA|    5|   100.0|
| Nongshim| Pack| South Korea| 4.25|    85.0|
| Nissin| Bowl| Japan| 4.5|    90.0|
| Nissin| Pack| Hong Kong|    5|   100.0|
| KOKA| Cup| Singapore| 3.5|    70.0|
| TRDP| Pack| India| 3.75|    75.0|
| Yamachan| Pack| USA|    5|   100.0|
| Binh Tay| Pack| Vietnam|    4|    80.0|
| Paldo| Pack| South Korea| 4|    80.0|
+-----+-----+-----+-----+
only showing top 20 rows
```

## SQL OPERATIONS

- Temp View

The Dataframe can be saved as temporary view which is present as long as that spark session is active

```
08 ✓ df.createOrReplaceTempView('Temp')
```

Above view can be used to perform Spark SQL queries

```
09 ✓ spark.sql("select * from Temp").show()
```

Brand	Style	Country	Ratings	Percentage
New Touch	Cup	Japan	3.75	75.0
Just Way	Pack	Taiwan	1	20.0
Nissin	Cup	USA	2.25	45.0
Wei Lih	Pack	Taiwan	2.75	55.0
Ching's Secret	Pack	India	3.75	75.0
Samyang Foods	Pack	South Korea	4.75	95.0
Acecook	Cup	Japan	4	80.0
Ikeda Shoku	Tray	Japan	3.75	75.0
Ripe'n'Dry	Pack	Japan	0.25	5.0
KOKA	Pack	Singapore	2.5	50.0
Tao Kae Noi	Pack	Thailand	5	100.0
Yamachan	Pack	USA	5	100.0
Nongshim	Pack	South Korea	4.25	85.0
Nissin	Bowl	Japan	4.5	90.0
Nissin	Pack	Hong Kong	5	100.0
KOKA	Cup	Singapore	3.5	70.0
TRDP	Pack	India	3.75	75.0
Yamachan	Pack	USA	5	100.0
Binh Tay	Pack	Vietnam	4	80.0
Paldo	Pack	South Korea	4	80.0

only showing top 20 rows

- Get Max Percentage from the dataset

```
spark.sql("select max(Percentage) as max_Percentage from Temp").show()
```

```
+-----+  
|max_Percentage|  
+-----+  
|          100.0|  
+-----+
```

- **Droping Percentage**

```
[35] df.drop('Percentage').show()
```

```
+-----+-----+-----+-----+  
|      Brand|style|   Country|Ratings|  
+-----+-----+-----+-----+  
| New Touch| Cup|    Japan|  3.75|  
| Just Way| Pack| Taiwan|    1|  
|   Nissin| Cup|     USA|  2.25|  
|   Wei Lih| Pack| Taiwan|  2.75|  
| Ching's Secret| Pack| India|  3.75|  
| Samyang Foods| Pack| South Korea| 4.75|  
| Acecook| Cup|    Japan|    4|  
| Ikeda Shoku| Tray|    Japan|  3.75|  
| Ripe'n'Dry| Pack|    Japan|  0.25|  
|       KOKA| Pack| Singapore|  2.5|  
| Tao Kae Noi| Pack| Thailand|    5|  
| Yamachan| Pack|     USA|    5|  
| Nongshim| Pack| South Korea| 4.25|  
|   Nissin| Bowl|    Japan|  4.5|  
|   Nissin| Pack| Hong Kong|    5|  
|       KOKA| Cup| Singapore|  3.5|  
|      TRDP| Pack| India|  3.75|  
| Yamachan| Pack|     USA|    5|  
|   Binh Tay| Pack| Vietnam|    4|  
|    Paldo| Pack| South Korea|    4|  
+-----+-----+-----+-----+  
only showing top 20 rows
```

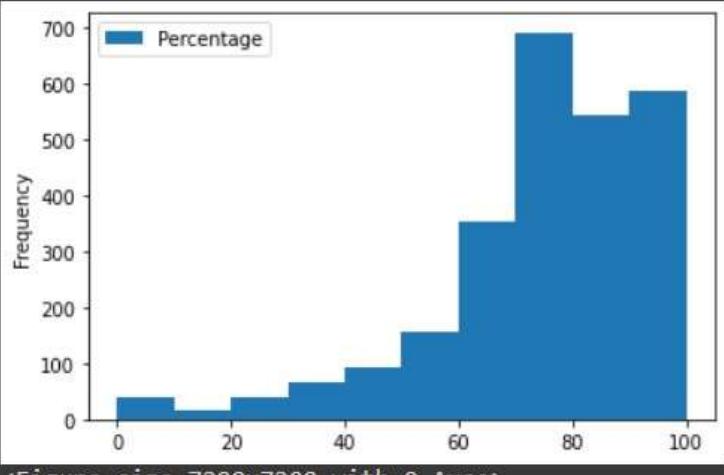
- **Using Pandas for plotting DataFrames:**

It converts the PySpark DataFrame into a Pandas DataFrame. After conversion, it's easy to create charts from pandas DataFrames using matplotlib or seaborn plotting tools.

```
[54] df_pandas = df.toPandas()
```

```
[56] import matplotlib.pyplot as plt
```

```
df_pandas.plot(kind='hist',x="style",y="Percentage")
plt.figure(figsize=(100, 100))
plt.show(20)
```



<Figure size 7200x7200 with 0 Axes>

# Practical 3

**Aim :** Implementation of hive

Hadoop : Hadoop is an open-source framework to store and process Big Data in a distributed environment. It contains two modules, one is MapReduce and another is Hadoop Distributed File System (HDFS).

1. MapReduce

2. HDFS

The Hadoop ecosystem contains different sub-projects (tools) such as Sqoop, Pig, and Hive that are used to help Hadoop modules

1. Sqoop

2. Pig

3.Hive

Step 1: Java-Installation

For hive installation first we have to verify java and Hadoop in our system

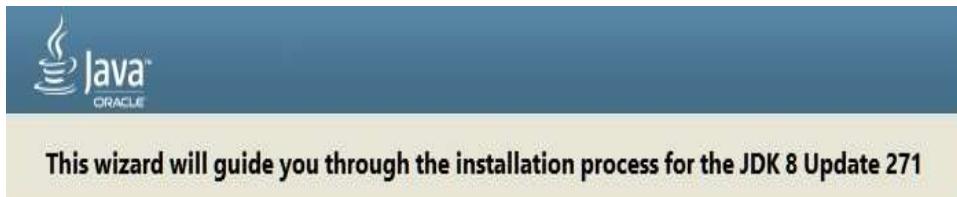
- Download the 8 version of java

Step 1: Run the Downloaded File Double-click the downloaded file to start the installation

Solaris SPARC 64-bit	88.75 MB	jdk-8u271-solaris-sparcv9.tar.gz
Solaris x64 (SVR4 package)	134.42 MB	jdk-8u271-solaris-x64.tar.Z
Solaris x64	92.52 MB	jdk-8u271-solaris-x64.tar.gz
Windows x86	154.48 MB	jdk-8u271-windows-i586.exe
Windows x64	166.79 MB	jdk-8u271-windows-x64.exe

NOTE: You will be required to create an Oracle Account to start Java 8 download of the file.

Step 2: Once the Java JDK 8 download is complete, run the exe for install JDK. Click Next



The terms under which this version of the software is licensed have changed.  
[Updated License Agreement](#)

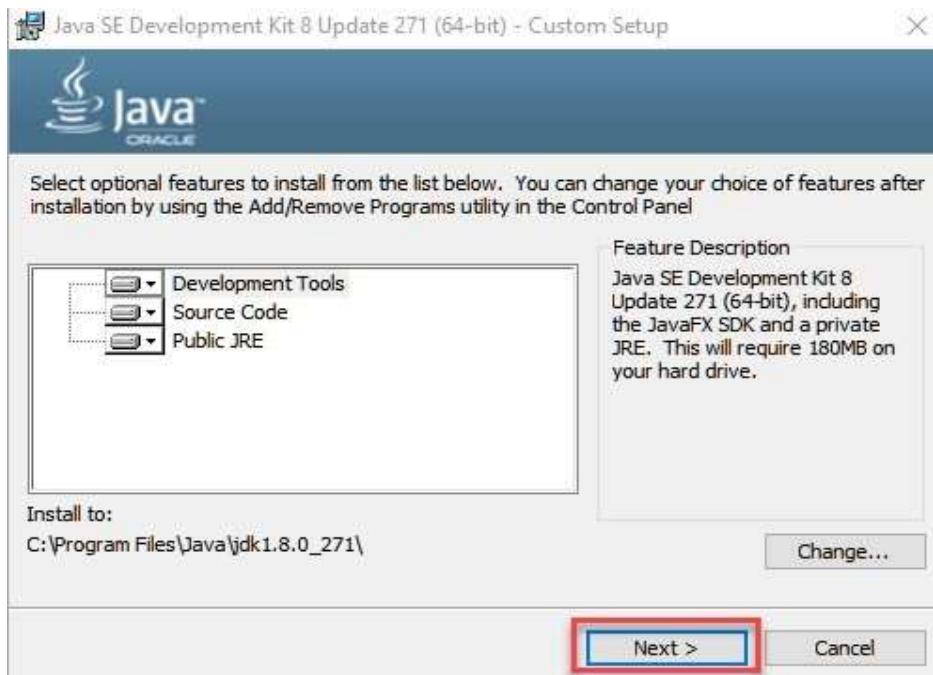
This version of the JDK no longer includes a copy of Java Mission Control (JMC). JMC is now available as a separate download.

Please visit <https://www.oracle.com/javase/jmc> for more information

No personal information is gathered as part of our install process.  
[Details on the information we collect](#)



Step 3: Select the PATH to install Java in Windows... You can leave it Default. Click next.

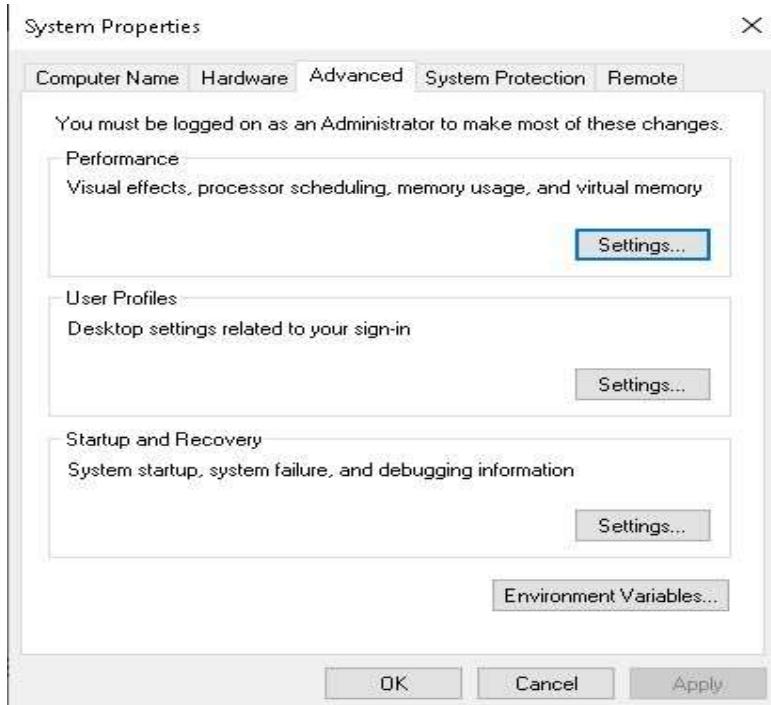


NOTE: Follow the onscreen instructions in succeeding steps to install Java 8 on Windows 10.

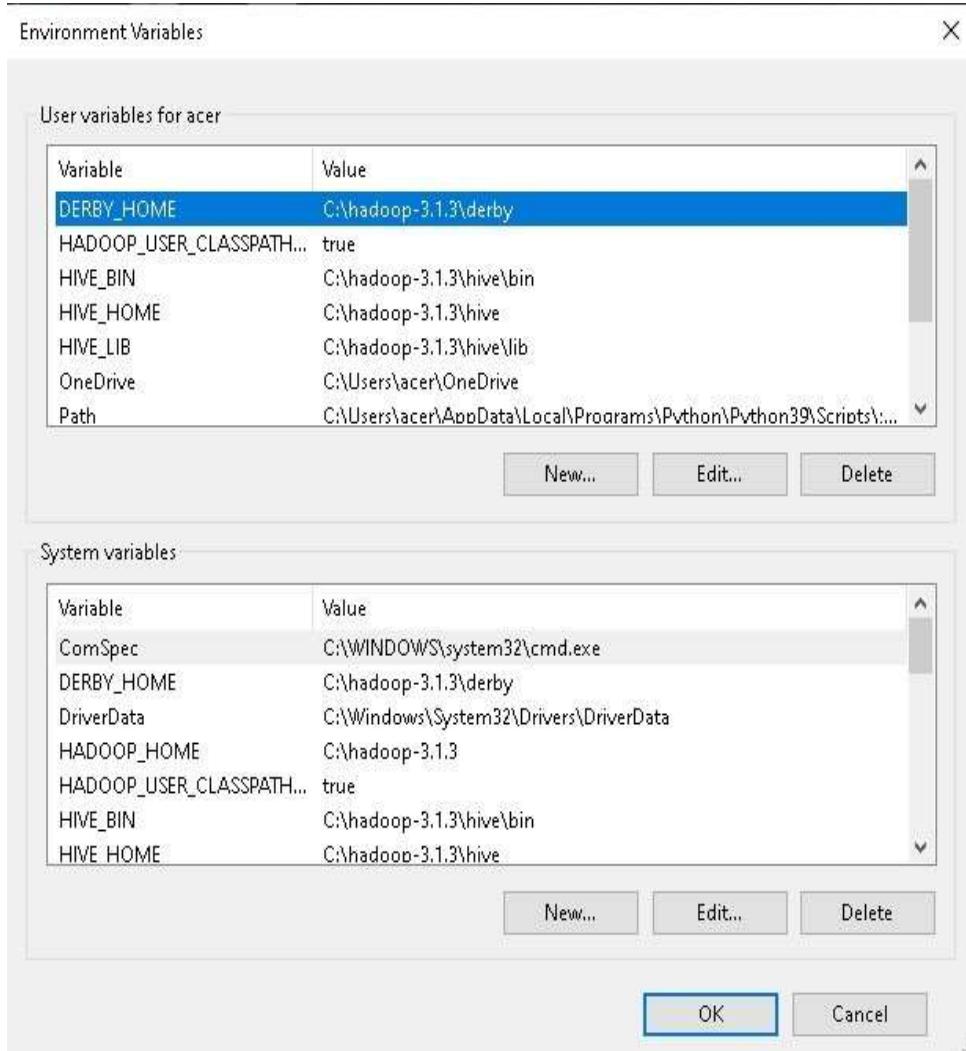
Step 4: Once you install Java in windows, click Close



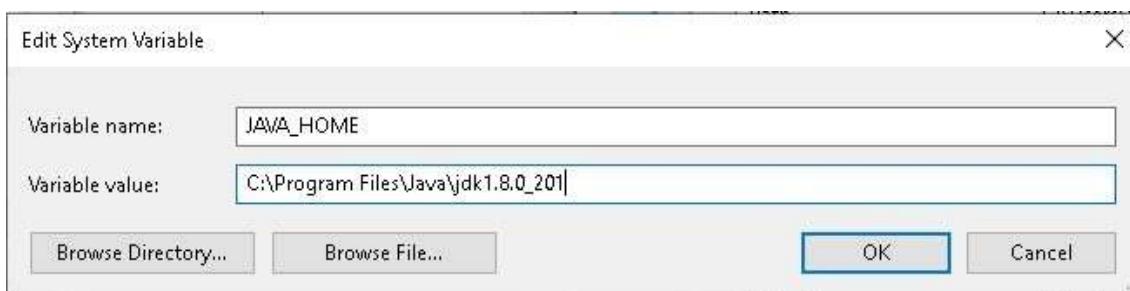
#### Step 5: Click on Environment Variables to set Java runtime environment



#### Step 6: Click on new Button of User variables/system variables



Step 7: Type JAVA\_HOME in the Variable name.



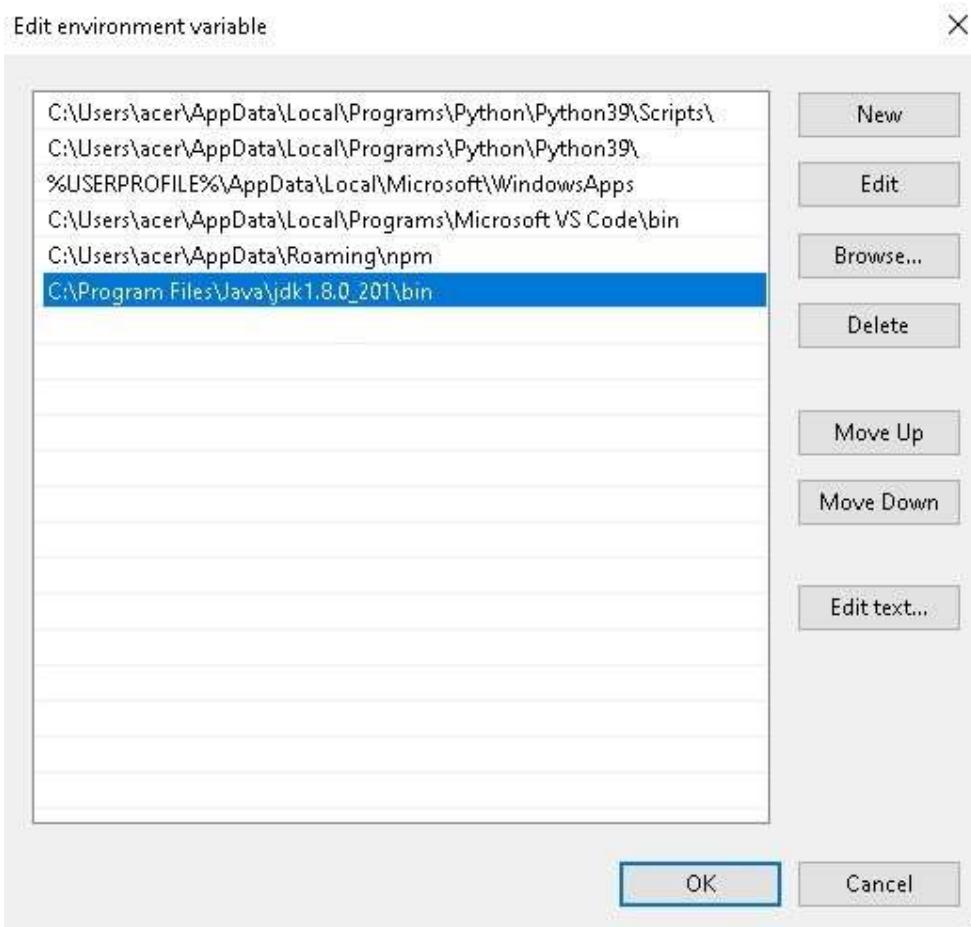
- Step 8 : Go to PATH

### User variables for acer

Variable	Value
HADOOP_USER_CLASSPATH...	true
HIVE_BIN	C:\hadoop-3.1.3\hive\bin
HIVE_HOME	C:\hadoop-3.1.3\hive
HIVE_LIB	C:\hadoop-3.1.3\hive\lib
OneDrive	C:\Users\acer\OneDrive
Path	C:\Users\acer\AppData\Local\Programs\Python\Python39\Scripts\... C:\Users\acer\... C:\Users\acer\AppData\Local\Temp
TEMP	C:\Users\acer\AppData\Local\Temp

New... Edit... Delete

- Step 9: go to edit and add the path of java jdk 8



Step 10: Go to command prompt and type “javac”and “java –version “

```
C:\Users\acer>javac
Usage: javac <options> <source files>
where possible options include:
  -g                                     Generate all debugging info
  -g:none                                Generate no debugging info
  -g:{lines,vars,source}                   Generate only some debugging info
  -nowarn                                Generate no warnings
  -verbose                               Output messages about what the compiler is doing
  -deprecation                           Output source locations where deprecated APIs are used
  -classpath <path>                      Specify where to find user class files and annotations
  -cp <path>                             Specify where to find user class files and annotations
  -sourcepath <path>                     Specify where to find input source files
  -bootclasspath <path>                  Override location of bootstrap class files
  -extdirs <dirs>                        Override location of installed extensions
  -endorseddirs <dirs>                  Override location of endorsed standards path
  -proc:{none,only}                      Control whether annotation processing and/or compilation is enabled
  -processor <class1>[,<class2>,<class3>...] Names of the annotation processors to be run
  -processorpath <path>                  Specify where to find annotation processors
  -parameters                            Generate metadata for reflection on method parameters
  -d <directory>                         Specify where to place generated class files
  -s <directory>                         Specify where to place generated source files
  -h <directory>                         Specify where to place generated native header files
  -implicit:{none,class}                Specify whether or not to generate class files for interfaces
  -encoding <encoding>                  Specify character encoding used by source files
  -source <release>                      Provide source compatibility with specified release
  -target <release>                     Generate class files for specific VM version
```

```
C:\Users\acer>java -version
java version "1.8.0_201"
Java(TM) SE Runtime Environment (build 1.8.0_201-b09)
Java HotSpot(TM) 64-Bit Server VM (build 25.201-b09, mixed mode)
```

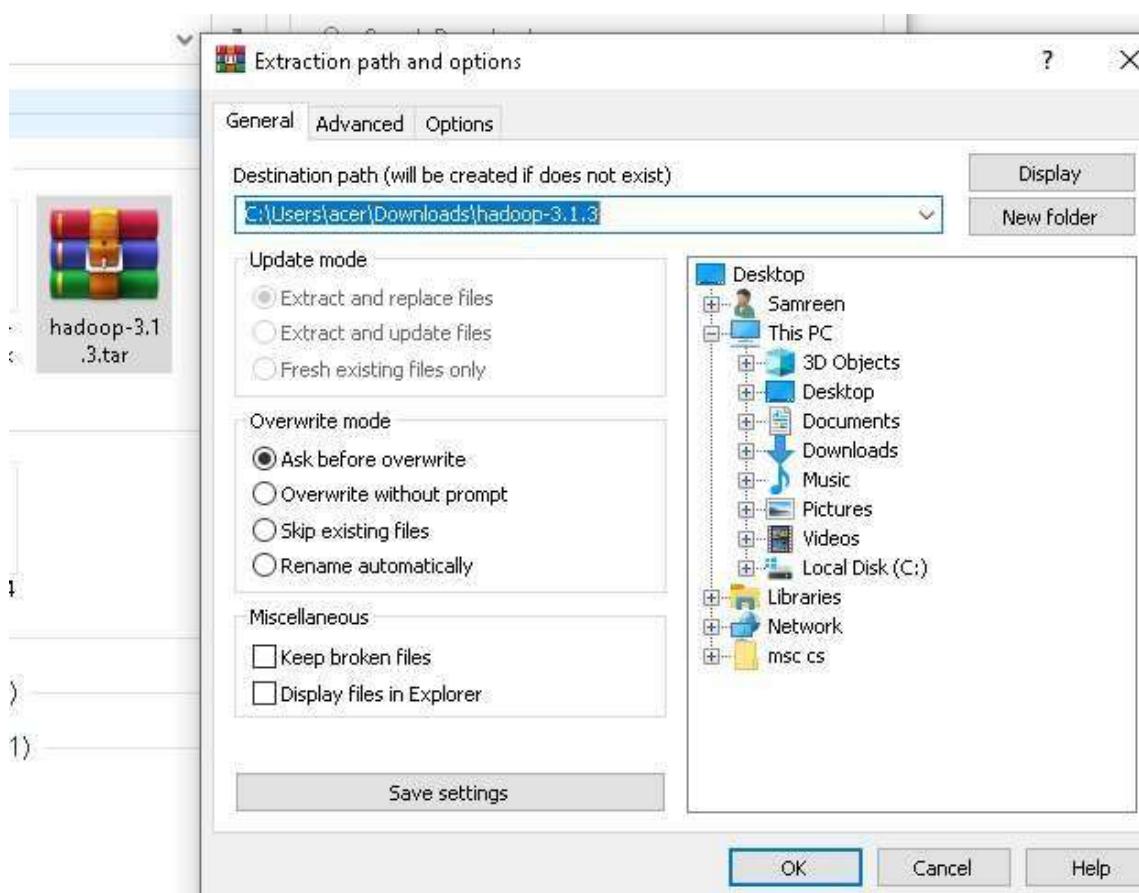
## Step 2:Installing Hadoop 3.2.1

Step 1: Download Hadoop tar file of Hadoop 3.2.1

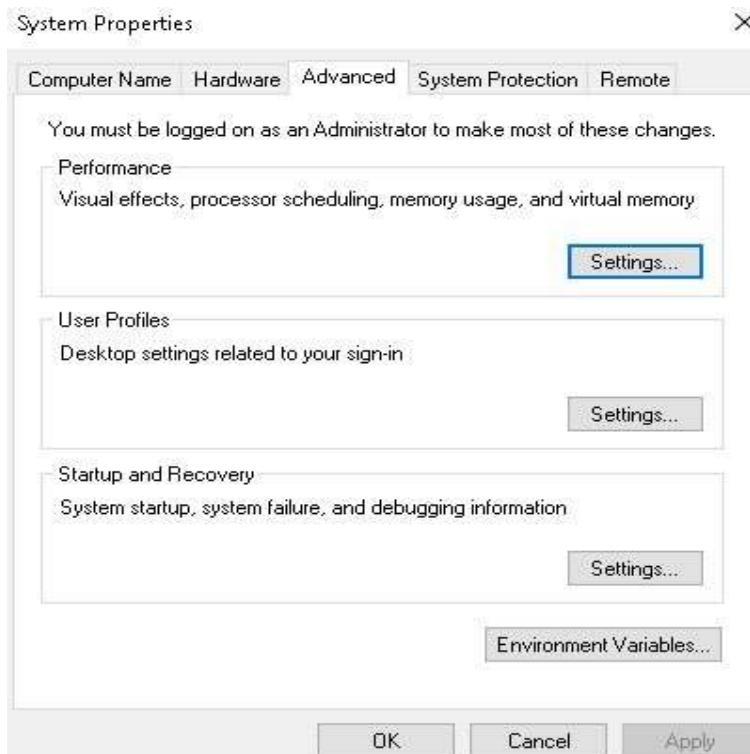
## Index of /dist/hadoop/common/hadoop-3.1.

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 Parent Directory		-	
 <a href="#">CHANGES.md</a>	2020-07-03 04:36	50K	
 <a href="#">CHANGES.md.asc</a>	2020-07-03 04:37	473	
 <a href="#">CHANGES.md.sha512</a>	2020-07-03 04:36	182	
 <a href="#">RELEASENOTES.md</a>	2020-07-03 04:36	2.8K	
 <a href="#">RELEASENOTES.md.asc</a>	2020-07-03 04:36	473	
 <a href="#">RELEASENOTES.md.sha512</a>	2020-07-03 04:37	187	
 <a href="#">hadoop-3.1.3-rat.txt</a>	2020-07-03 04:36	1.7M	
 <a href="#">hadoop-3.1.3-rat.txt.asc</a>	2020-07-03 04:36	473	
 <a href="#">hadoop-3.1.3-rat.txt.sha512</a>	2020-07-03 04:37	192	
 <a href="#">hadoop-3.1.3-site.tar.gz</a>	2020-07-03 04:36	38M	
 <a href="#">hadoop-3.1.3-site.tar.gz.asc</a>	2020-07-03 04:36	473	
 <a href="#">hadoop-3.1.3-site.tar.gz.sha512</a>	2020-07-03 04:36	196	
 <a href="#">hadoop-3.1.3-src.tar.gz</a>	2020-07-03 04:37	28M	

Step 2: extract hadoop file in c drive



Step 3: after extracting file go to system and environment



Step 4 : Add a path of Hadoop in environment

System variables	
Variable	Value
OS	Windows_NT
Path	%DERBY_HOME%\bin;C:\Program Files (x86)\VMware\VMware Pla...
PATHEXT	.COM;.EXE;.BAT;.CMD;.VBS;.VBE;.JS;.JSE;.WSF;.WSH;.MSC
PROCESSOR_ARCHITECTURE	AMD64
PROCESSOR_IDENTIFIER	AMD64 Family 21 Model 112 Stepping 0, AuthenticAMD
PROCESSOR_LEVEL	21
PROCESSOR_REVISION	7000

New... Edit... Delete

OK Cancel

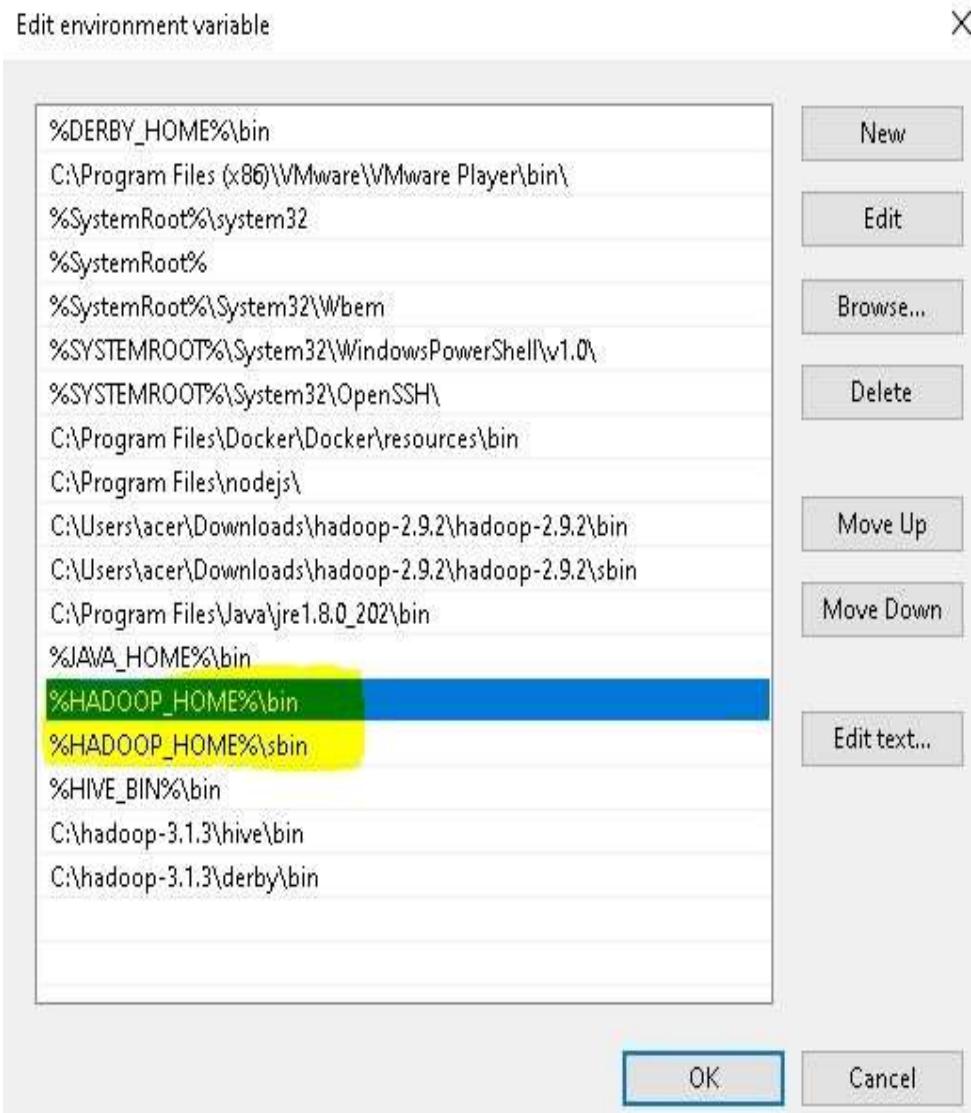
Note : Hadoop bin and sbin both path should added in environment variable

Edit System Variable X

Variable name: HADOOP\_HOME

Variable value: C:\hadoop-3.1.3

Browse Directory... Browse File... OK Cancel



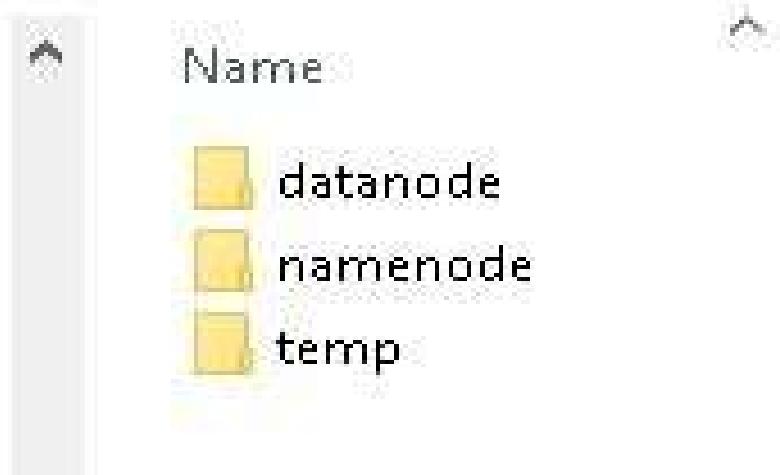
#### Step 5:Configuring Hadoop cluster

1. %HADOOP\_HOME%\etc\hadoop\hdfs-site.xml
2. %HADOOP\_HOME%\etc\hadoop\core-site.xml
3. %HADOOP\_HOME%\etc\hadoop\mapred-site.xml
4. %HADOOP\_HOME%\etc\hadoop\yarn-site.xml

#### Step 6: Create a directory to store all master node

- 1 namenode
- 2 datanode
- 3 temp

<< hadoop-3.1.3 >> data



Step 7: Open “C:\hadoop-3.1.3\etc\Hadoop\hdfs-site.xml” file location and add the following properties

```
<configuration>
<property>
<name>dfs.namenode.name.dir</name>
<value>file:///opt/hdfs/namenode</value>
<description>NameNode directory for namespace and transaction logs storage.</description>
</property>
<property>
<name>dfs.datanode.data.dir</name>
<value>file:///opt/hdfs/datanode</value>
<description>DataNode directory</description>
</property>
<property>
<name>dfs.replication</name>
<value>3</value>
</property>
<property>
<name>dfs.permissions</name>
<value>false</value>
</property>
<property>
<name>dfs.datanode.use.datanode.hostname</name>
<value>false</value>
</property>
<property>
<name>dfs.namenode.datanode.registration.ip-hostname-check</name>
<value>false</value>
</property>
</configuration>
```

Step 8: Open “C:\hadoop-3.1.3\etc\Hadoop\core-site.xml” file location and add the following properties

```
<configuration>
<property>
<name>fs.defaultFS</name>
<value>hdfs://hadoop-namenode:9820/</value>
<description>NameNode URI</description>
</property>
<property>
<name>io.file.buffer.size</name>
<value>131072</value>
<description>Buffer size</description>
</property>
</configuration>
```

Step 9: Open “C:\hadoop-3.1.3\etc\Hadoop\yarn-site.xml” file location and add the following properties



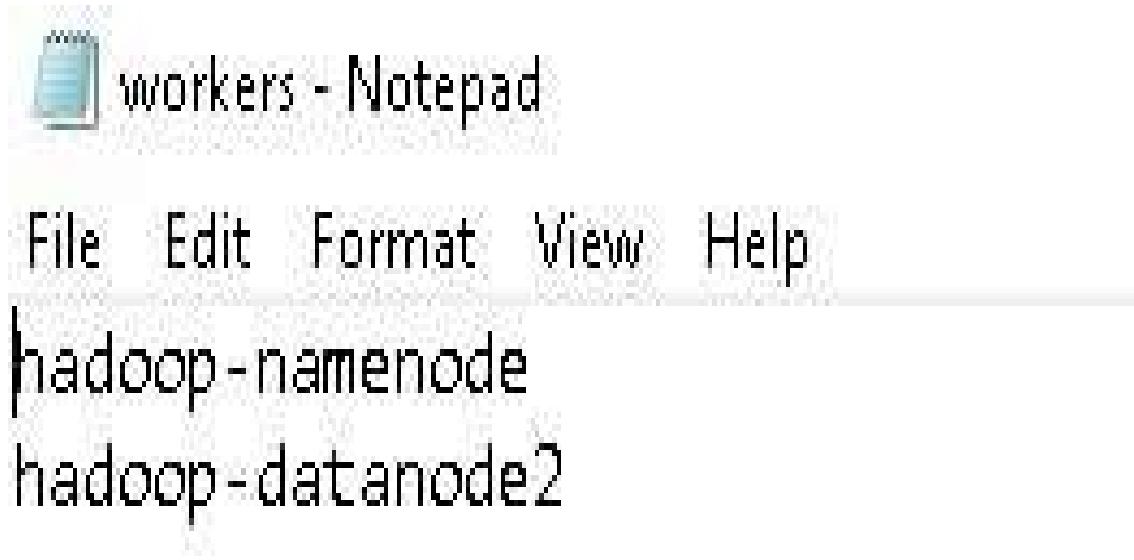
The screenshot shows a Notepad window titled “\*Untitled - Notepad”. The menu bar includes File, Edit, Format, View, and Help. The main content area contains XML configuration code for the YARN service.

```
*Untitled - Notepad
File Edit Format View Help
<configuration>
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
<description>Yarn Node Manager Aux Service</description>
</property>
<property>
<name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
<value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property>
<name>yarn.nodemanager.local-dirs</name>
<value>file:///opt/yarn/local</value>
</property>
<property>
<name>yarn.nodemanager.log-dirs</name>
<value>file:///opt/yarn/logs</value>
</property>
</configuration>
```

Step 10: Open “C:\hadoop-3.1.3\etc\Hadoop\mapred-site.xml” file location and add the following properties

```
<configuration>
<property>
<name>mapreduce.framework.name</name>
<value>yarn</value>
<description>MapReduce framework name</description>
</property>
<property>
<name>mapreduce.jobhistory.address</name>
<value>hadoop-namenode:10020</value>
<description>Default port is 10020.</description>
</property>
<property>
<name>mapreduce.jobhistory.webapp.address</name>
<value> hadoop-namenode:19888</value>
<description>Default port is 19888.</description>
</property>
<property>
<name>mapreduce.jobhistory.intermediate-done-dir</name>
<value>/mr-history/tmp</value>
<description>Directory where history files are written by MapReduce jobs.</description>
</property>
<property>
<name>mapreduce.jobhistory.done-dir</name>
<value>/mr-history/done</value>
<description>Directory where history files are managed by the MR JobHistory Server.</description>
</property>
</configuration>
```

Step 11: Add you namenode and data node in workers file C:\hadoop-3.1.3\etc\Hadoop\workers



Step 12: Go to cmd run as administration and type command “startall”



The screenshot shows a browser window with the URL `hadoop-namenode:9870/dfshealth.html#tab-overview`. The page has a green header bar with tabs: Hadoop (selected), Overview, Datanodes, Datanode Volume Failures, Snapshot, Startup Progress, and Utilities. The main content area is titled "Overview 'hadoop-namenode:9820' (active)". Below the title is a table with the following data:

<b>Started:</b>	Mon Dec 05 02:25:07 +0530 2022
<b>Version:</b>	3.1.3, rba631c436b806728f8ec2f54ab1e289526c90579
<b>Compiled:</b>	Thu Sep 12 08:17:00 +0530 2019 by ztang from branch-3.1.3
<b>Cluster ID:</b>	CID-470dbb83-51da-468b-814e-c4f55656f162
<b>Block Pool ID:</b>	BP-113919451-172.25.16.1-1669760656575

### Step 3: Hive – Installation

Step 1 : downloading hive tar file

## Index of /dist/hive/hive-2.1.0

Name	Last modified	Size	Description
<a href="#">Parent Directory</a>		-	
<a href="#"> apache-hive-2.1.0-bin.tar.gz</a>	2016-06-21 01:26	143M	
<a href="#"> apache-hive-2.1.0-bin.tar.gz.asc</a>	2016-06-21 01:26	819	
<a href="#"> apache-hive-2.1.0-bin.tar.gz.md5</a>	2016-06-21 01:26	70	
<a href="#"> apache-hive-2.1.0-src.tar.gz</a>	2016-06-21 01:26	18M	
<a href="#"> apache-hive-2.1.0-src.tar.gz.asc</a>	2016-06-21 01:26	819	
<a href="#"> apache-hive-2.1.0-src.tar.gz.md5</a>	2016-06-21 01:26	70	

Step 2 : Download derby file after extracting copy a lib file of derby and paste it in hive lib file

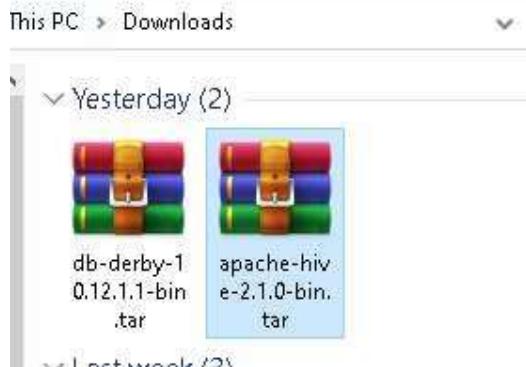
# Index of /dist/db/derby/db-derby-10.12.1.1

<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 <a href="#">Parent Directory</a>		-	
 <a href="#">db-derby-10.12.1.1-bin.tar.gz</a>	2015-10-10 14:38	18M	
 <a href="#">db-derby-10.12.1.1-bin.tar.gz.asc</a>	2015-10-10 14:38	194	
 <a href="#">db-derby-10.12.1.1-bin.tar.gz.md5</a>	2015-10-10 14:38	33	
 <a href="#">db-derby-10.12.1.1-bin.zip</a>	2015-10-10 14:38	20M	

Step 3 : Download the hive-xml file and after extracting the file this file copy paste in hive conf file location

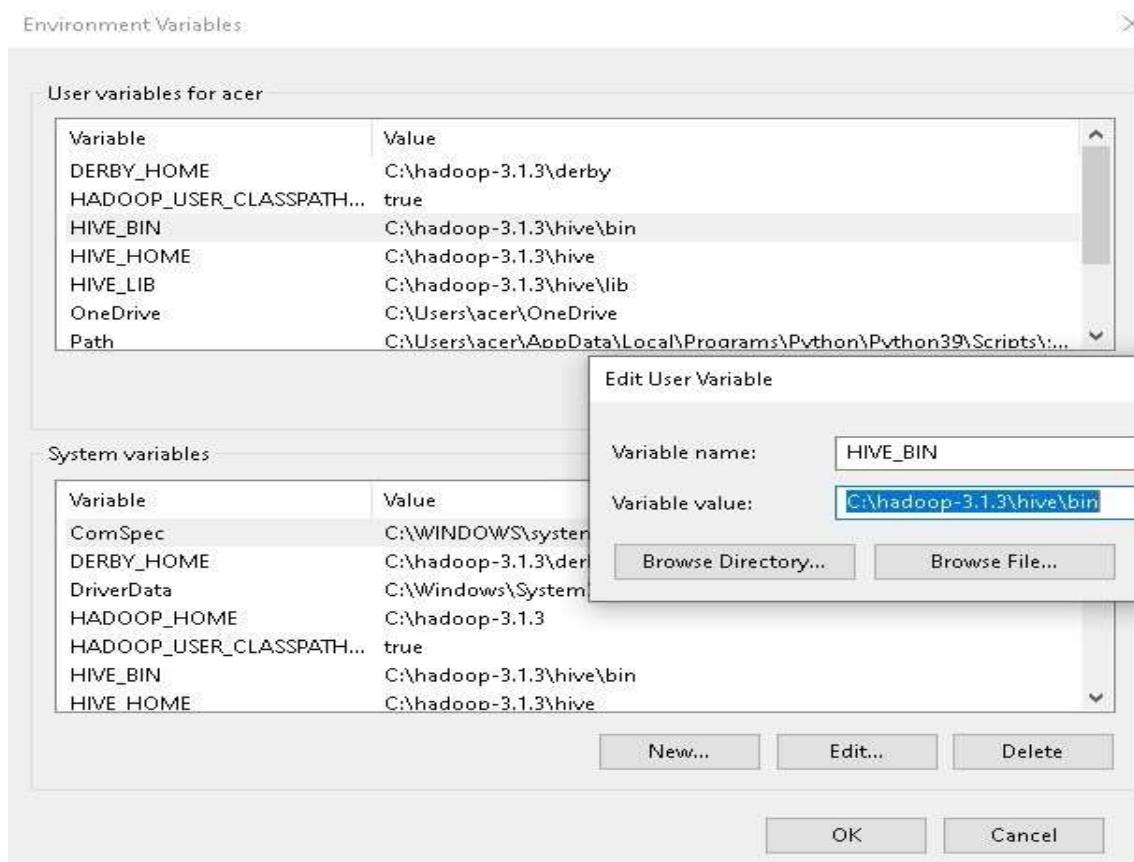
```
<?xml version="1.0"?>
<?xmlstylesheet type="text/xsl" href="configuration.xsl"?>
<configuration><property> <name>javax.jdo.option.ConnectionURL</name>
<value>jdbc:derby://localhost:1527/metastore_db;create=true</value>
<description>JDBC connect string for a JDBC metastore</description>
</property><property>
<name>javax.jdo.option.ConnectionDriverName</name>
<value>org.apache.derby.jdbc.ClientDriver</value>
<description>Driver class name for a JDBC metastore</description>
</property>
<property>
<name>hive.server2.enable.impersonation</name>
<description>Enable user impersonation for HiveServer2</description>
<value>true</value>
</property>
<property>
<name>hive.server2.authentication</name>
<value>NONE</value>
<description> Client authentication types. NONE: no authentication check LDAP: LDAP/AD based authentication KERBEROS: Kerberos/GSSAPI authentication CUSTOM: Custom authentication provider (Use with property hive.server2.custom.authentication.class) </description>
</property>
<property>
<name>datanucleus.autoCreateTables</name>
<value>True</value>
</property>
</configuration>
```

#### Step 4: Extract the hive tar file in Hadoop file



Name		Date modified
bin		29-09-2021 01:42
data		30-11-2022 03:10
derby		04-12-2022 08:31
etc		30-11-2022 03:03
hive		04-12-2022 08:31
include		30-11-2022 03:09
lib		30-11-2022 03:03
libexec		30-11-2022 03:03
local		05-12-2022 02:25
logs		05-12-2022 02:25
sbin		30-11-2022 03:03
share		30-11-2022 03:04
LICENSE		04-09-2019 03:01
NOTICE		04-09-2019 03:01
README		04-09-2019 03:01

### Step 5 :Set the path in environment variable



### Edit User Variable

Variable name:

Variable value:

### Edit User Variable

Variable name:

Variable value:

### Edit User Variable

Variable name:

Variable value:

### Edit System Variable

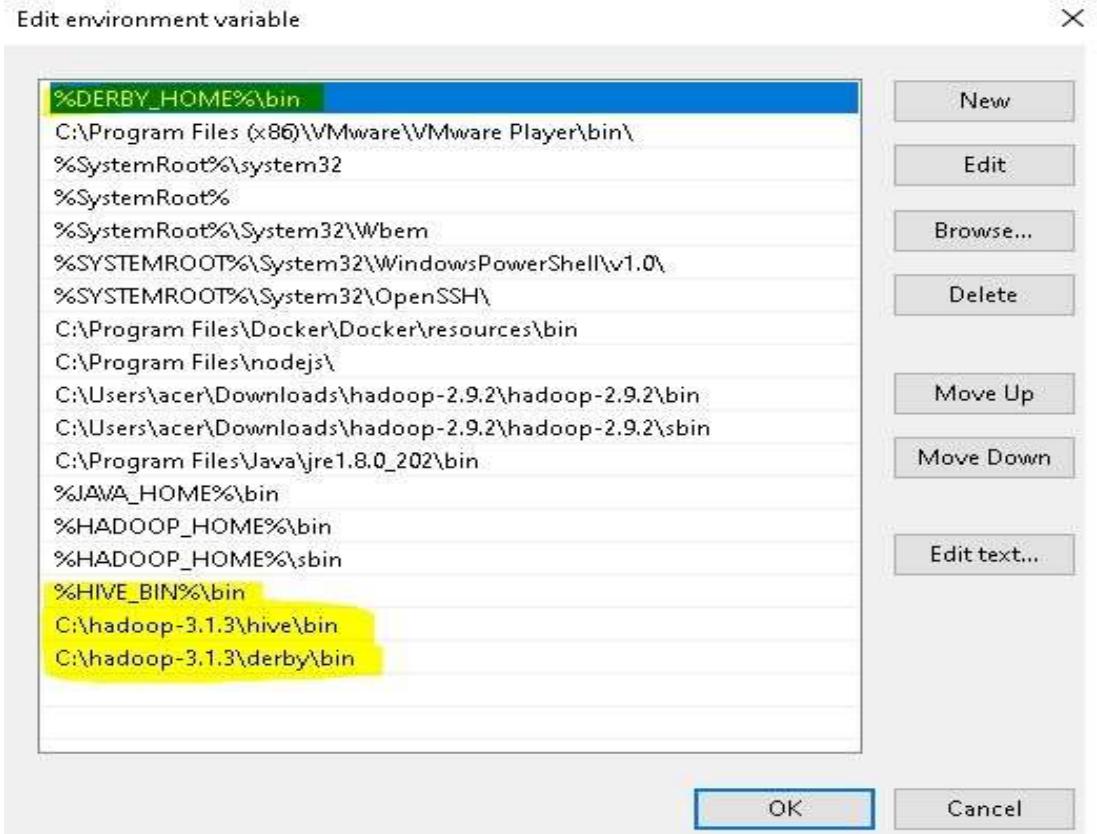
Variable name:

Variable value:

### Edit System Variable

Variable name:

Variable value:



## Step 6: First start Hadoop by using command ‘start-dfs’

```

Administrator: Command Prompt
Microsoft Windows [Version 10.0.19045.2251]
(c) Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>cd C:\hadoop-3.1.3\sbin

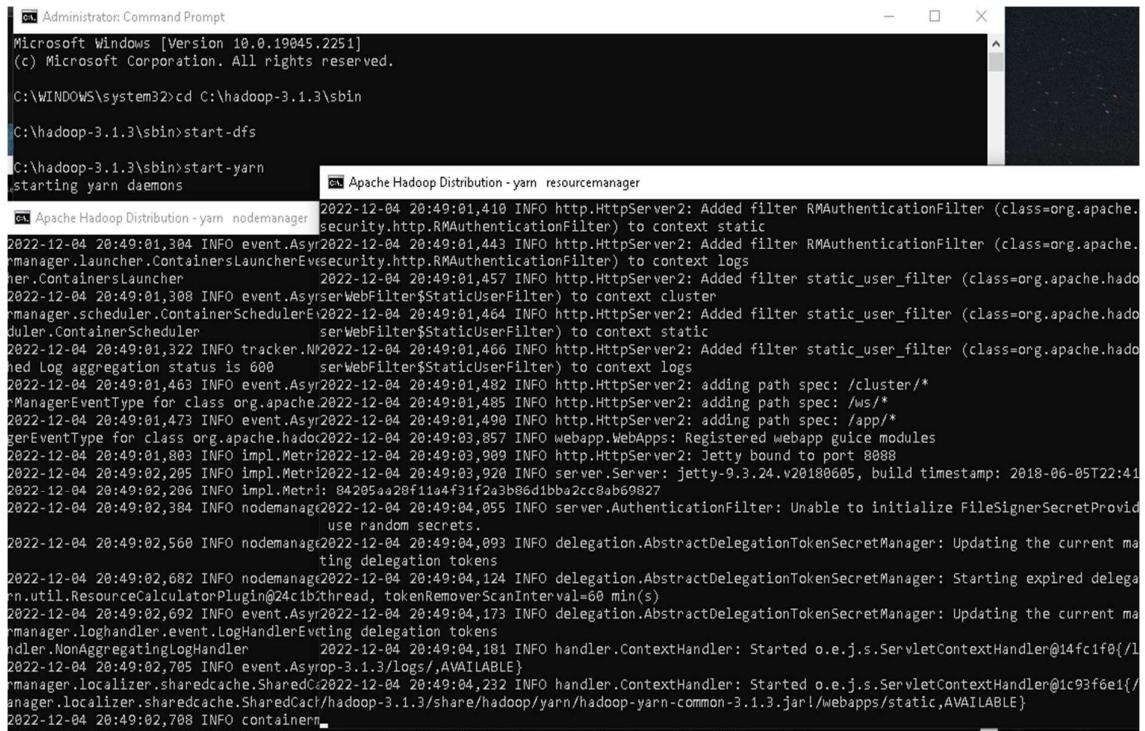
C:\hadoop-3.1.3\sbin>start-dfs

C:\hadoop-3.1.3\sbin>start-yarn
Apache Hadoop Distribution - hadoop namenode

2022-12-04 20:47:56,062 INFO blo Apache Hadoop Distribution - hadoop datanode
2022-12-04 20:47:56,064 INFO blo 2022-12-04 20:47:58,706 INFO impl.FsDatasetImpl: Scanning block pool BP-113919451-172.25.16.1-166976065657
-replicated blocks completed in \hadoop-3.1.3\data\datanode...
2022-12-04 20:47:55,998 INFO nam 2022-12-04 20:47:58,706 INFO impl.FsDatasetImpl: Time taken to scan block pool BP-113919451-172.25.16.1-166976065657
2022-12-04 20:47:56,085 INFO nam C:\hadoop-3.1.3\data\datanode: 104ms
2022-12-04 20:47:56,086 INFO nam 2022-12-04 20:47:58,706 INFO impl.FsDatasetImpl: Total time to scan all replicas for block pool BP-113919451-172.25.16.1-166976065657: 119ms
2022-12-04 20:47:56,082 INFO ipc 2022-12-04 20:47:58,712 INFO impl.FsDatasetImpl: Adding replicas to map for block pool BP-113919451-172.25.16.1-166976065657
2022-12-04 20:47:56,036 INFO ipc 2022-12-04 20:47:58,712 INFO impl.FsDatasetImpl: Adding replicas to map for block pool BP-113919451-172.25.16.1-166976065657 on volume C:\hadoop-3.1.3\data\datanode...
2022-12-04 20:47:56,262 INFO nam 2022-12-04 20:47:58,715 INFO impl.BlockPoolSlice: Replica Cache file: C:\hadoop-3.1.3\data\datanode\current
name space=2 2022-12-04 20:47:58,715 INFO impl.BlockPoolSlice: Replica Cache file: C:\hadoop-3.1.3\data\datanode\current\replicas doesn't exist
storage space=0 -172.25.16.1-166976065657\current\replicas
storage types=RAM_DISK=0, SSD=0, 2022-12-04 20:47:58,733 INFO impl.FsDatasetImpl: Time to add replicas to map for block pool BP-113919451-172.25.16.1-166976065657
2022-12-04 20:47:56,344 INFO blo 9760656575 on volume C:\hadoop-3.1.3\data\datanode: 18ms
2022-12-04 20:47:56,344 INFO blo 2022-12-04 20:47:58,736 INFO impl.FsDatasetImpl: Total time to add all replicas to map for block pool BP-113919451-172.25.16.1-166976065657: 00 milliseconds
2022-12-04 20:47:59,159 INFO hdfs 5.16.1-1669760656575: 26ms
deuid=e7524046f-6fc3-4edc-9831-a2022-12-04 20:47:58,736 INFO datanode.VolumeScanner: VolumeScanner(C:\hadoop-3.1.3\data\datanode, DS-e127b70dbb83-51da-468b-814e-c4f556f-a987-de957b56fd30): no suitable block pools found to scan. Waiting 1409760256 ms.
2022-12-04 20:47:59,167 INFO net 2022-12-04 20:47:58,956 INFO datanode.DirectoryScanner: Periodic Directory Tree Verification scan starting
2022-12-04 20:47:59,169 INFO blo 41 AM with interval of 2160000ms
2022-12-04 20:47:59,169 INFO blo 2022-12-04 20:47:59,028 INFO datanode.DataNode: Block pool BP-113919451-172.25.16.1-1669760656575 (Datanode (127.0.0.1:9866)).
2022-12-04 20:48:00,104 INFO blo f-6fc3-4edc-9831-a13224be6c20 service to hadoop-namenode/127.0.0.1:9820 beginning handshake with NN
56fd30 for DN 127.0.0.1:9866 2022-12-04 20:47:59,480 INFO datanode.DataNode: Block pool Block pool BP-113919451-172.25.16.1-1669760656575
2022-12-04 20:48:00,255 INFO Blo 7524046f-6fc3-4edc-9831-a13224be6c20 service to hadoop-namenode/127.0.0.1:9820 successfully registered
for DS-e127ba84-fcb9-4286-a987-d2022-12-04 20:47:59,486 INFO datanode.DataNode: For namenode hadoop-namenode/127.0.0.1:9820 using BLOCKREP
2022-12-04 20:48:00,264 INFO Blo f 2160000msec CACHEREPORT_INTERVAL of 10000msec Initial delay: 0msec; heartBeatInterval=3000
86-a987-de957b56fd30 node Datanode 2022-12-04 20:48:00,402 INFO datanode.DataNode: Successfully sent block report 0xb503d433e1025edd, containing 9864, infoSecurePort=0, ipcPort report(s), of which we sent 1. The reports had 0 total blocks and used 1 RPC(s). This took 13 msec to generate
c=1669760656575), blocks: 0, has 0 mssecs for RPC and NN processing. Got back one command: FinalizeCommand/5.
2022-12-04 20:48:00,407 INFO datanode.DataNode: Got finalize command for block pool BP-113919451-172.25.16.1-1669760656575
75

```

## Step 7: Then start yarn by using command ‘start-yarn’ in cmd

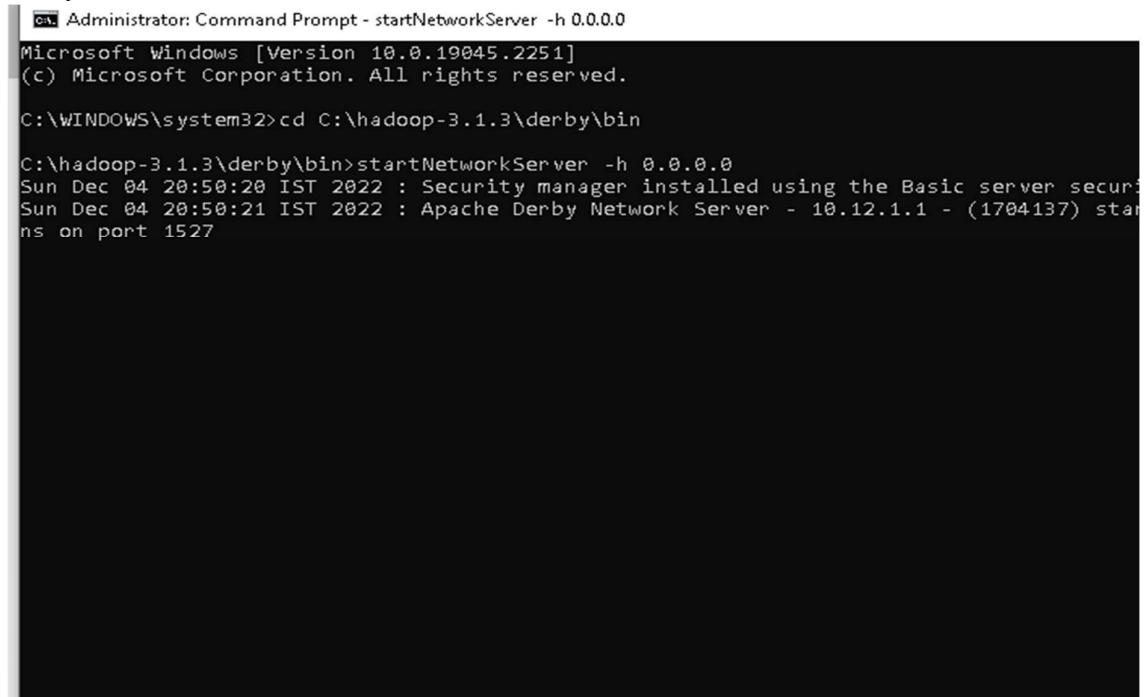


```
Administrator: Command Prompt
Microsoft Windows [Version 10.0.19045.2251]
(c) Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>cd C:\hadoop-3.1.3\sbin
C:\hadoop-3.1.3\sbin>start-dfs
C:\hadoop-3.1.3\sbin>start-yarn
starting yarn daemons
Apache Hadoop Distribution - yarn resourcemanager
2022-12-04 20:49:01,410 INFO http.HttpServer2: Added filter RMAuthenticationFilter (class=org.apache.hadoop.security.http.RMAuthenticationFilter) to context static
2022-12-04 20:49:01,304 INFO event.Asyrmgr 2022-12-04 20:49:01,443 INFO http.HttpServer2: Added filter RMAuthenticationFilter (class=org.apache.hadoop.yarn.launcher.ContainersLauncherEventSecurity.http.RMAuthenticationFilter) to context logs
2022-12-04 20:49:01,308 INFO event.Asyrmgr 2022-12-04 20:49:01,457 INFO http.HttpServer2: Added filter static_user_filter (class=org.apache.hadoop.yarn.scheduler.ContainerSchedulerEvent$StaticUserFilter) to context cluster
2022-12-04 20:49:01,322 INFO tracker.Ni 2022-12-04 20:49:01,466 INFO http.HttpServer2: Added filter static_user_filter (class=org.apache.hadoop.yarn.log.aggregation.LogAggregationStatus is 600) serWebFilter$StaticUserFilter) to context logs
2022-12-04 20:49:01,463 INFO event.Asyrmgr 2022-12-04 20:49:01,482 INFO http.HttpServer2: adding path spec: /cluster/*
2022-12-04 20:49:01,473 INFO event.Asyrmgr 2022-12-04 20:49:01,485 INFO http.HttpServer2: adding path spec: /ws/*
2022-12-04 20:49:01,482 INFO event.Asyrmgr 2022-12-04 20:49:01,490 INFO http.HttpServer2: adding path spec: /app/*
2022-12-04 20:49:01,492 INFO event.Asyrmgr 2022-12-04 20:49:01,497 INFO http.HttpServer2: adding path spec: /logs/*
2022-12-04 20:49:01,499 INFO event.Asyrmgr 2022-12-04 20:49:01,500 INFO http.HttpServer2: Jetty bound to port 8088
2022-12-04 20:49:02,205 INFO impl.Metric 2022-12-04 20:49:03,920 INFO server.Server: Jetty-9.3.24.v20180605, build timestamp: 2018-06-05T22:41:22Z, build version: v20180605, build os: Linux, build arch: amd64, host os: Windows 10, host arch: amd64, host ip: 192.168.1.11
2022-12-04 20:49:02,206 INFO impl.Metric: 84205a28f11a4f31f2a3b86d1ba2cc8ab69827
2022-12-04 20:49:02,384 INFO nodemanag 2022-12-04 20:49:04,055 INFO server.AuthenticationFilter: Unable to initialize FileSignerSecretProvider. Using random secrets.
2022-12-04 20:49:02,560 INFO nodemanag 2022-12-04 20:49:04,093 INFO delegation.AbstractDelegationTokenSecretManager: Updating the current mapping delegation tokens
2022-12-04 20:49:02,682 INFO nodemanag 2022-12-04 20:49:04,124 INFO delegation.AbstractDelegationTokenSecretManager: Starting expired delegation token removal
2022-12-04 20:49:02,692 INFO event.Asyrmgr 2022-12-04 20:49:04,173 INFO delegation.AbstractDelegationTokenSecretManager: Updating the current manager log handler event LogHandlerEvting delegation tokens
2022-12-04 20:49:02,693 INFO event.Asyrmgr 2022-12-04 20:49:04,181 INFO handler.ContextHandler: Started o.e.j.s.ServletContextHandler@14fc1f0{/l}
2022-12-04 20:49:02,705 INFO event.Asyrmgr 2022-12-04 20:49:04,232 INFO handler.ContextHandler: Started o.e.j.s.ServletContextHandler@1c93f6e1{/l}
2022-12-04 20:49:02,708 INFO container
```

Step 8: Open new cmd and change the directory by derby bin

"c:\hadoop\derby\bin" run command "startNetworkServer -h 0.0.0.0



```
Administrator: Command Prompt - startNetworkServer -h 0.0.0.0
Microsoft Windows [Version 10.0.19045.2251]
(c) Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>cd C:\hadoop-3.1.3\derby\bin
C:\hadoop-3.1.3\derby\bin>startNetworkServer -h 0.0.0.0
Sun Dec 04 20:50:20 IST 2022 : Security manager installed using the Basic server security
Sun Dec 04 20:50:21 IST 2022 : Apache Derby Network Server - 10.12.1.1 - (1704137) starts on port 1527
```

Step 9: Open new cmd and change directory by hive bin and run command "hive"

```
Administrator: Command Prompt
Microsoft Windows [Version 10.0.19045.2251]
(c) Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>cd C:\hadoop-3.1.3\hive\bin

C:\hadoop-3.1.3\hive\bin>hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/hadoop-3.1.3/hive/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/hadoop-3.1.3/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/loggerFactory]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
ERROR StatusLogger No log4j2 configuration file found. Using default configuration: logging only errors to the console.
Connecting to jdbc:hive2:///
com.google.common.base.Preconditions.checkNotNull(ZLjava/lang/String;Ljava/lang/Object;)V
Beeline version 2.1.0 by Apache Hive
com.google.common.base.Preconditions.checkNotNull(ZLjava/lang/String;Ljava/lang/Object;)V
Connection is already closed.

C:\hadoop-3.1.3\hive\bin>
```

## Step 10: Hive as installed Now we can use hive shell

```
Microsoft Windows [Version 10.0.15063]
(c) 2017 Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>cd C:\hive\bin

C:\hive\bin>hive
ERROR StatusLogger No log4j2 configuration file found. Using default configuration: logging only errors to the console.
Connecting to jdbc:hive2:///
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/C:/hive/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/C:/hadoop1/hadoop-2.3.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.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]
Connected to: Apache Hive (version 2.1.0)
Driver: Hive JDBC (version 2.1.0)
Transaction isolation: TRANSACTION_REPEATABLE_READ
Beeline version 2.1.0 by Apache Hive
hive> CREATE DATABASE a;
OK
No rows affected (2.06 seconds)
hive> SHOW DATABASES;
OK
a
default
2 rows selected (0.901 seconds)
hive>
```

## Step 11: Commands in Hive

- 1.Create : It is used to create a table or database

```
hive> create database IF NOT EXISTS college;
```

```
OK
```

```
Time taken: 0.062 seconds
```

2. Show : It is used to show Database, Table, Properties, etc

```
hive> Show databases;
OK
default
firstdb
Time taken: 0.039 seconds, Fetched: 2 row(s)
```

3. ALTER : It is used to make changes to the existing table

Command ALTER TABLE [db\_name].old\_table\_name RENAME  
TO [db\_name].new\_table\_name;

```
hive> ALTER TABLE college.senior_students RENAME TO college.college_students;
OK
Time taken: 0.247 seconds
```

4. DESCRIBE: It describes the table columns

Command DESCRIBE [db\_name.] table\_name[.col\_name ( [.field\_name])]

```
hive> describe college.college_students;
OK
id          bigint          unique id for each student
name        string          student name
age          int            student age between 16-26
fee          double         student college fee
city         string         cities to which students belong
state        string         student home address streets
zip          bigint         student address zip code
Time taken: 0.064 seconds, Fetched: 7 row(s)
```

5. TRUNCATE:Used to permanently truncate and delete the rows of table

Command “ truncate table [db\_name].table\_name”

```
hive> truncate table college.senior_students;
OK
Time taken: 0.141 seconds
```

## 6. Create : It use to create table in hive warehouse location

```
hive> create table txnrecords(txnno INT, txndate STRING, custno INT, amount DOUBLE,category STRING, product STRING, city STRING, state STRING, spendby STRING) row format delimited fields terminated by ',' stored as textfile;
OK
Time taken: 1.163 seconds
hive>
```

## 7 Load : Load operation is used to move the data into corresponding Hive table.

```
hive> LOAD DATA INPATH '/txns1.txt' OVERWRITE INTO TABLE txnrecords;
Loading data to table retail.txnrecords
Deleted hdfs://localhost/user/hive/warehouse/retail.db/txnrecords
OK
Time taken: 0.263 seconds
hive>
```

The screenshot shows a web browser interface with the following details:

- Address bar: HDFS:/user/hive/warehouse/r...
- URL: /user/hive/warehouse/retail.db/txnrrecords/txns1.txt
- Content area:
  - Goto: /user/hive/warehouse/retail... go
  - Links: Go back to dir listing, Advanced view/download options, View Next chunk
  - Data preview:

```
00000000.06-25-2011,4007024,040.33,Exercise & Fitness,Cardio Machine Accessories,Clarksville,Tennessee,credit
00000001.05-26-2011,4006742,198.44,Exercise & Fitness,Weightlifting Gloves,Long Beach,California,credit
00000002.06-01-2011,4009775,005.58,Exercise & Fitness,Weightlifting Machine Accessories,Anaheim,California,credit
00000003.06-05-2011,4002199,198.19,Gymnastics,Gymnastics Rings,Milwaukee,Wisconsin,credit
00000004.12-17-2011,4002613,698.81,Team Sports,Field Hockey,Nashville ,Tennessee,credit
00000005.02-14-2011,4007591,193.63,Outdoor Recreation,Camping & Backpacking & Hiking,Chicago,Illinois,credit
00000006.10-28-2011,4002190,027.89,Puzzles,Jigsaw Puzzles,Charleston,South Carolina,credit
00000007.07-14-2011,4002964,696.01,Outdoor Play Equipment,Sandboxes,Columbus,Ohio,credit
00000008.01-17-2011,4007361,010.44,Winter Sports,Snowmobiling,Des Moines,Iowa,credit
00000009.05-17-2011,4004798,152.46,Jumping,Bungee Jumping,St. Petersburg,Florida,credit
```

## 8 Count : Count aggregate function is used count the total number of the records in a table.

```
hive> select count(*) from txnrecords;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_201402270420_0005, Tracking URL = http://localhost:50030/jobdetails.jsp?jobid=job_201402270420_0005
Kill Command = /usr/lib/hadoop/bin/hadoop job -Dmapred.job.tracker=localhost:8021 -kill job_201402270420_0005
2014-02-28 20:02:41,231 Stage-1 map = 0%,  reduce = 0%
2014-02-28 20:02:48,293 Stage-1 map = 50%,  reduce = 0%
2014-02-28 20:02:49,389 Stage-1 map = 100%,  reduce = 0%
2014-02-28 20:02:55,350 Stage-1 map = 100%,  reduce = 33%
2014-02-28 20:02:56,367 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_201402270420_0005
OK
50000
Time taken: 19.027 seconds
```

## Practical 4

**Aim:** Implementation of HBASE

### What is Hadoop?

Apache Hadoop is an open-source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyse massive datasets in parallel more quickly.

Hadoop consists of four main modules:

1. Hadoop Distributed File System (HDFS) – A distributed file system that runs on standard or low-end hardware. HDFS provides better data throughput than traditional file systems, in addition to high fault tolerance and native support of large datasets.
2. Yet Another Resource Negotiator (YARN) – Manages and monitors cluster nodes and resource usage. It schedules jobs and tasks.
3. MapReduce – A framework that helps programs do the parallel computation on data. The map task takes input data and converts it into a dataset that can be computed in key value pairs. The output of the map task is consumed by reduce tasks to aggregate output and provide the desired result.
4. Hadoop Common – Provides common Java libraries that can be used across all modules.

### What is HBase?

Apache HBase is an open-source, NoSQL, distributed big data store. It enables random, strictly consistent, real-time access to petabytes of data. HBase is very effective for handling large, sparse datasets.

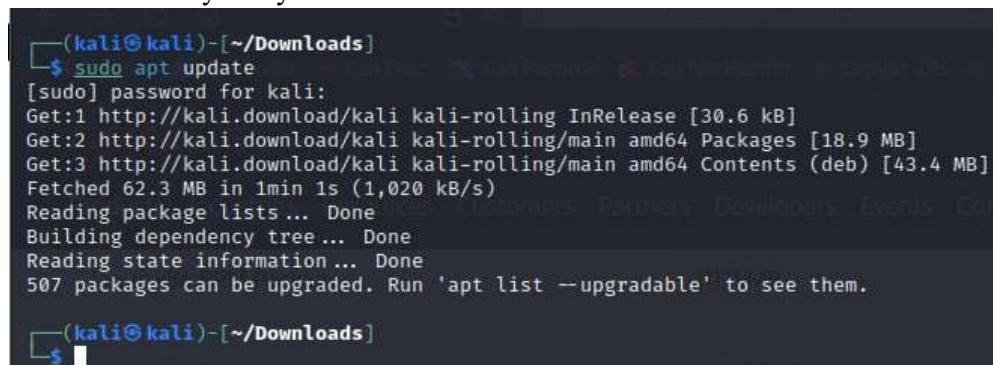
HBase integrates seamlessly with Apache Hadoop and the Hadoop ecosystem and runs on top of the Hadoop Distributed File System (HDFS) or Amazon S3 using Amazon Elastic MapReduce (EMR) file system, or EMRFS. HBase serves as a direct input and output to the Apache MapReduce framework for Hadoop, and works with Apache Phoenix to enable SQL-like queries over HBase tables.

### Installation:

#### A. Java 8:

(Note: in order to run Hadoop, we will require Java 8, i.e., jdk1.8.0\_341)

1. Update the Linux distribution on your system



```
(kali㉿kali)-[~/Downloads]
$ sudo apt update
[sudo] password for kali:
Get:1 http://kali.download/kali kali-rolling InRelease [30.6 kB]
Get:2 http://kali.download/kali kali-rolling/main amd64 Packages [18.9 MB]
Get:3 http://kali.download/kali kali-rolling/main amd64 Contents (deb) [43.4 MB]
Fetched 62.3 MB in 1min 1s (1,020 kB/s)
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
507 packages can be upgraded. Run 'apt list --upgradable' to see them.

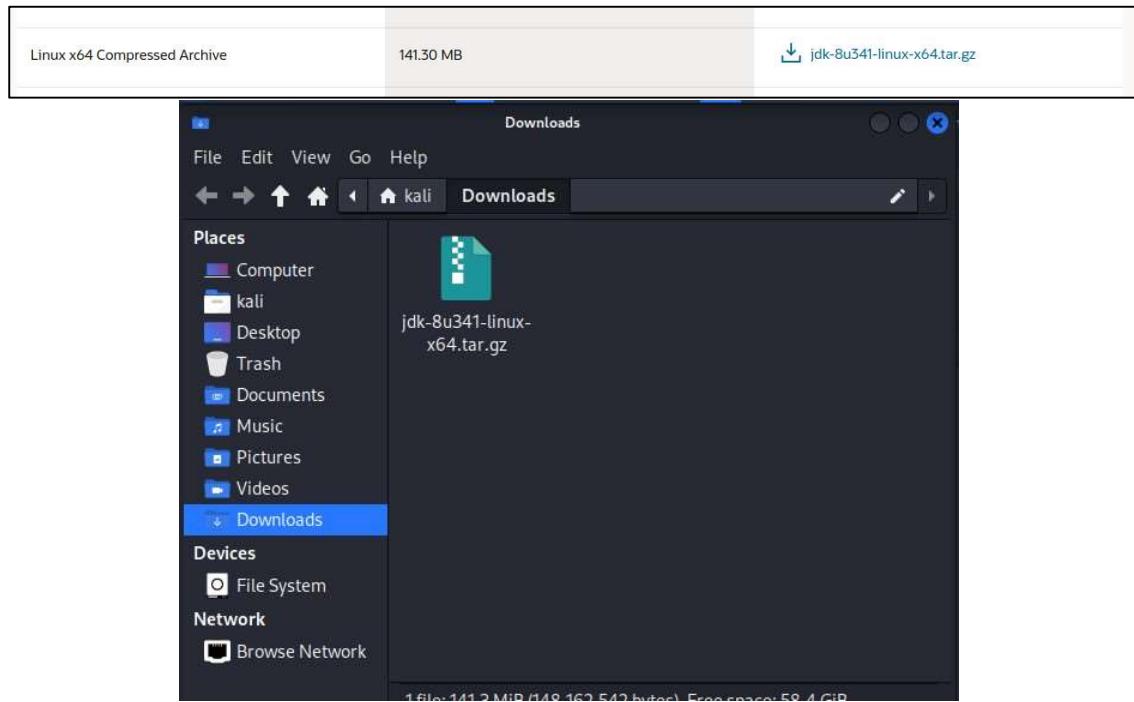
(kali㉿kali)-[~/Downloads]
$
```

2. Update your Java Development Kit (JDK)

```
(kali㉿kali)-[~/Downloads]
$ sudo apt-get install default-jdk
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following packages were automatically installed and are no longer required:
libatk1.0-data libev4 libexpat-tiny-perl libflac8 libfm8 libhttp-server-simple-perl libilmbase25 liblberc3 liblist-moreutils-perl
liblist-moreutils-xs-perl libopenexr25 libopenh264-6 libperl5.34 libplacebo192 libpoppler118 libpython3.9-minimal libpython3.9-stdlib libsvtav1enc0
libwebsockets16 libwireshark15 libwiretap12 libwsutil13 openjdk-11-jre perl-modules-5.34 python3-dataclasses-json python3-limiter
python3-marshmallow-enum python3-mypy-extensions python3-responses python3-spyse python3-token-bucket python3-typing-inspect python3.9
python3.9-minimal
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
default-jdk-headless default-jre default-jre-headless java-common libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libxau-dev libxcb1-dev
libxdmcp-dev libxt-dev openjdk-17-jdk openjdk-17-jdk-headless openjdk-17-jre openjdk-17-jre-headless x11proto-dev xorg-sgml-doctools xtrans-dev
Suggested packages:
libice-doc libsm-doc libx11-doc libxcb-doc libxt-doc openjdk-17-demo openjdk-17-source visualvm fonts-ipafont-gothic fonts-ipafont-mincho
fonts-way-microhei | fonts-way-zenhei fonts-indic
The following NEW packages will be installed:
default-jdk default-jdk-headless libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libxau-dev libxcb1-dev libxdmcp-dev libxt-dev openjdk-17-jdk
openjdk-17-jdk-headless openjdk-17-jre openjdk-17-jre-headless x11proto-dev xorg-sgml-doctools xtrans-dev
The following packages will be upgraded:
default-jre default-jre-headless java-common
3 upgraded, 17 newly installed, 0 to remove and 504 not upgraded.
Need to get 280 MB of archives.
After this operation, 442 MB of additional disk space will be used.
Do you want to continue? [Y/n] Y
Get:1 http://kali.download/kali kali-rolling/main amd64 java-common all 0.73 [6,268 B]
Get:2 http://http.kali.org/kali kali-rolling/main amd64 openjdk-17-jre-headless amd64 17.0.5+8-2 [43.6 MB]
Get:3 http://kali.download/kali kali-rolling/main amd64 default-jre amd64 2:1.17-73 [1,060 B]
Get:4 http://kali.download/kali kali-rolling/main amd64 default-jre-headless amd64 2:1.17-73 [2,812 B]
```

### 3. Download Java from the given link:

<https://www.oracle.com/java/technologies/javase/javase8u211-later-archivedownloads.html>



As seen in the image below, the Java version is 17.

Path “/usr/lib/jvm” is where all the java files are available.

```
(kali㉿kali)-[~]
└─$ java --version
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
openjdk 17.0.5 2022-10-18
OpenJDK Runtime Environment (build 17.0.5+8-Debian-2)
OpenJDK 64-Bit Server VM (build 17.0.5+8-Debian-2, mixed mode, sharing)

(kali㉿kali)-[~]
└─$ cd /usr/lib/jvm/
└─(kali㉿kali)-[/usr/lib/jvm]
└─$ ll
total 12
lrwxrwxrwx 1 root root 25 Oct 31 14:07 default-java → java-1.17.0-openjdk-amd64
lrwxrwxrwx 1 root root 21 Jul 20 04:54 java-1.11.0-openjdk-amd64 → java-11-openjdk-amd64
lrwxrwxrwx 1 root root 21 Oct 19 10:23 java-1.17.0-openjdk-amd64 → java-17-openjdk-amd64
drwxr-xr-x 7 root root 4096 Aug 8 06:05 java-11-openjdk-amd64
drwxr-xr-x 9 root root 4096 Dec 1 03:01 java-17-openjdk-amd64
drwxr-xr-x 2 root root 4096 Dec 1 03:02 openjdk-17

(kali㉿kali)-[/usr/lib/jvm]
└─$
```

#### 4. Extract Java 8 files in the path “/usr/lib/jvm”.

```
(kali㉿kali)-[/usr/lib/jvm]
└─$ sudo tar -xvzf ~/Downloads/jdk-8u341-linux-x64.tar.gz
[sudo] password for kali:
[jdk1.8.0_341/COPYRIGHT
[jdk1.8.0_341/LICENSE
[jdk1.8.0_341/README.html
[jdk1.8.0_341/THIRDPARTYLICENSEREADME.txt
[jdk1.8.0_341/bin/java-rmi.cgi
[jdk1.8.0_341/bin/appletviewer
[jdk1.8.0_341/bin/extcheck
[jdk1.8.0_341/bin/idlj
[jdk1.8.0_341/bin/jar
[jdk1.8.0_341/bin/jarsigner
[jdk1.8.0_341/bin/javac
```

As seen in the image below files are extract in the folder “jdk1.8.0\_341”.

```
(kali㉿kali)-[/usr/lib/jvm]
└─$ ll
total 16
lrwxrwxrwx 1 root root 25 Oct 31 14:07 default-java → java-1.17.0-openjdk-amd64
lrwxrwxrwx 1 root root 21 Jul 20 04:54 java-1.11.0-openjdk-amd64 → java-11-openjdk-amd64
lrwxrwxrwx 1 root root 21 Oct 19 10:23 java-1.17.0-openjdk-amd64 → java-17-openjdk-amd64
drwxrwxrwx 7 root root 4096 Aug 8 06:05 java-11-openjdk-amd64
drwxrwxrwx 9 root root 4096 Dec 1 03:01 java-17-openjdk-amd64
drwxrwxrwx 8 root root 4096 Dec 1 12:10 jdk1.8.0_341
drwxrwxrwx 2 root root 4096 Dec 1 03:02 openjdk-17
```

#### 5. Add the files path in the environment file as shown in the image.

We need to add

- “/usr/lib/jvm/jdk1.8.0\_341/bin”
- “/usr/lib/jvm/jdk1.8.0\_341/jre/bin”

```

[(kali㉿kali)-[~/usr/lib/jvm]
$ cd jdk1.8.0_341
[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ vi /etc/environment
[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ sudo vi /etc/environment
[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ sudo vi /etc/environment
[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ more /etc/environment
# START KALI-DEFAULTS CONFIG
# Everything from here and until STOP KALI-DEFAULTS CONFIG
# was installed by the kali-defaults package, and it will
# be removed if ever the kali-defaults package is removed.
# If you want to disable a line, please do NOT remove it,
# as it would be added back when kali-defaults is upgraded.
# Instead, comment the line out, and your change will be
# preserved across upgrades.
PATH=/usr/local/sbin:/usr/local/bin:/usr/bin:/sbin:/bin:/usr/local/games:/usr/games:/usr/lib/jvm/jdk1.8.0_341/bin:/usr/lib/jvm/jdk1.8.0_341/jre/bin
COMMAND_NOT_FOUND_INSTALL_PROMPT=1
POWERSHELL_UPDATECHECK=Off
POWERSHELL_TELEMETRY_OPTOUT=1
DOTNET_CLI_TELEMETRY_OPTOUT=1
# STOP KALI-DEFAULTS CONFIG

```

## 6. Now we have to update the Java version to the Java 8.

```

[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --install "/usr/bin/java" "java" "/usr/lib/jvm/jdk1.8.0_341/bin/java" 0
[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --install "/usr/bin/java" "java" "/usr/lib/jvm/jdk1.8.0_341/bin/javac" 0
[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --set java /usr/lib/jvm/jdk1.8.0_341/bin/java
update-alternatives: using /usr/lib/jvm/jdk1.8.0_341/bin/java to provide /usr/bin/java (java) in manual mode
[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --set java /usr/lib/jvm/jdk1.8.0_341/bin/javac
update-alternatives: using /usr/lib/jvm/jdk1.8.0_341/bin/javac to provide /usr/bin/java (java) in manual mode
[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ 

[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --config java
There are 4 choices for the alternative java (providing /usr/bin/java).

Selection    Path                      Priority    Status
-----    -----
0            /usr/lib/jvm/java-17-openjdk-amd64/bin/java    1711      auto mode
1            /usr/lib/jvm/java-11-openjdk-amd64/bin/java    1111      manual mode
2            /usr/lib/jvm/java-17-openjdk-amd64/bin/java    1711      manual mode
* 4           /usr/lib/jvm/jdk1.8.0_341/bin/java        0         manual mode
*           /usr/lib/jvm/jdk1.8.0_341/bin/javac        0         manual mode

Press <enter> to keep the current choice[*], or type selection number: 3
update-alternatives: using /usr/lib/jvm/jdk1.8.0_341/bin/java to provide /usr/bin/java (java) in manual mode
[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ java -version
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
java version "1.8.0_341"
Java(TM) SE Runtime Environment (build 1.8.0_341-b10)
Java HotSpot(TM) 64-Bit Server VM (build 25.341-b10, mixed mode)

[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ 

[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --install "/usr/bin/javac" "javac" "/usr/lib/jvm/jdk1.8.0_341/bin/javac" 1
[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --set javac /usr/lib/jvm/jdk1.8.0_341/bin/javac
update-alternatives: using /usr/lib/jvm/jdk1.8.0_341/bin/javac to provide /usr/bin/javac (javac) in manual mode
[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --config javac
There are 2 choices for the alternative javac (providing /usr/bin/javac).

Selection    Path                      Priority    Status
-----    -----
0            /usr/lib/jvm/java-17-openjdk-amd64/bin/javac    1711      auto mode
1            /usr/lib/jvm/java-17-openjdk-amd64/bin/javac    1711      manual mode
* 2           /usr/lib/jvm/jdk1.8.0_341/bin/javac        1         manual mode

Press <enter> to keep the current choice[*], or type selection number: 2
[(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
$ 

```

As shown in the below image, the Java version is changed to Java 8.

```
(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
└─$ java -version
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
java version "1.8.0_341"
Java(TM) SE Runtime Environment (build 1.8.0_341-b10)
Java HotSpot(TM) 64-Bit Server VM (build 25.341-b10, mixed mode)

(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
└─$ javac -version
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
javac 1.8.0_341

(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
└─$
```

## B. Hadoop:

1. Create a separate user for Hadoop. For installation as well as execution.

```
(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
└─$ sudo addgroup hadoop
Adding group `hadoop' (GID 1001) ...
Done.

(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
└─$ sudo adduser --ingroup hadoop dion
Adding user `dion' ...
Adding new user `dion' (1002) with group `hadoop (1001)' ...
Creating home directory `/home/dion' ...
Copying files from `/etc/skel' ...
New password:
Retype new password:
passwd: password updated successfully
Changing the user information for dion
Enter the new value, or press ENTER for the default
      Full Name []: Big Data Project Dion
      Room Number []:
      Work Phone []:
      Home Phone []:
      Other []:
Is the information correct? [Y/n] Y
Adding new user `dion' to supplemental / extra groups `users' ...
Adding user `dion' to group `users' ...

(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
└─$ sudo adduser dion sudo
Adding user `dion' to group `sudo' ...
Done.

(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
└─$
```

2. Now we will install the SSH Server.

This is necessary for us to start the various components in Hadoop.

```
(kali㉿kali)-[~/usr/lib/jvm/jdk1.8.0_341]
└─$ sudo apt install openssh-server openssh-client -y
Reading package lists...
Building dependency tree...
Reading state information...
openSSH-server is already the newest version (1:9.0p1-1+b2).
openSSH-server set to manually installed.
openSSH-client is already the newest version (1:9.0p1-1+b2).
The following packages were automatically installed and are no longer required:
libatk1.0-data libev4 libexporter-tiny-perl libflac8 libfmt8 libhttp-server-simple-perl libilmbase25 liblrc3 liblist-moreutils-perl
liblist-moreutils-xs-perl libopenexr25 libopenp264-6 libpopt15.34 libpoptable192 libpoppier118 libpython3.9-minimal libpython3.9-stdlib libsvtavenc0
libwebsocketst16 libwireshark15 libwiredtap12 libwsutil13 openjdk-11-jre perl-modules-5.34 python3-dataclasses-json python3-limiter
python3-marshall-enum python3-mypy-extensions python3-responses python3-spuse python3-token-bucket python3-typing-inspect python3.9
python3.9-minimal
Use 'sudo apt autoremove' to remove them.
0 upgraded, 0 newly installed, 0 to remove and 504 not upgraded.
```

3. Switch user from “kali” to “dion”.

```
└──(kali㉿kali)-[/usr/lib/jvm/jdk1.8.0_341]
└─$ su - dion
Password:
└──(dion㉿kali)-[~]
```

4. Create SSH key

```
└──(dion㉿kali)-[~]
└─$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/dion/.ssh/id_rsa):
Created directory '/home/dion/.ssh'.
Your identification has been saved in /home/dion/.ssh/id_rsa
Your public key has been saved in /home/dion/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:53VANRDTkAMZg1MSnuV+DqqAlfMJgGydrDbU8m4kp34 dion㉿kali
The key's randomart image is:
+---[RSA 3072]---+
|   O***Bo |
| .. . =.00 ... |
| .. =    0 .. .. |
| o + . . . . |
| = +=    S + o . |
| . Bo + . + = . |
| ..o. o . . . |
| ..E . . . |
| .. . . . |
+---[SHA256]---+
```

5. Save the Key.

```
└──(dion㉿kali)-[~]
└─$ cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
-bash: /home/dion/: Is a directory
```

6. Now, we will start a SSH server on localhost.

```

(dion㉿kali)-[~]
└─$ sudo service ssh status
[sudo] password for dion:
● ssh.service - OpenBSD Secure Shell server
  Loaded: loaded (/lib/systemd/system/ssh.service; disabled; preset: disabled)
  Active: inactive (dead)
    Docs: man:sshd(8)
          man:sshd_config(5)

(dion㉿kali)-[~]
└─$ sudo service ssh start

(dion㉿kali)-[~]
└─$ sudo service ssh status
● ssh.service - OpenBSD Secure Shell server
  Loaded: loaded (/lib/systemd/system/ssh.service; disabled; preset: disabled)
  Active: active (running) since Fri 2022-12-02 04:57:30 EST; 10s ago
    Docs: man:sshd(8)
          man:sshd_config(5)
  Process: 33683 ExecStartPre=/usr/sbin/sshd -t (code=exited, status=0/SUCCESS)
  Main PID: 33691 (sshd)
    Tasks: 1 (limit: 4619)
   Memory: 2.8M
      CPU: 2.735s
     CGroup: /system.slice/ssh.service
             └─33691 "sshd: /usr/sbin/sshd -D [listener] 0 of 10-100 startups"

Dec 02 04:57:27 kali systemd[1]: Starting OpenBSD Secure Shell server...
Dec 02 04:57:30 kali sshd[33691]: Server listening on 0.0.0.0 port 22.
Dec 02 04:57:30 kali sshd[33691]: Server listening on :: port 22.
Dec 02 04:57:30 kali systemd[1]: Started OpenBSD Secure Shell server.

(dion㉿kali)-[~]
└─$ █

(dion㉿kali)-[~]
└─$ ssh localhost
The authenticity of host 'localhost (::1)' can't be established.
ED25519 key fingerprint is SHA256:4d/a0Z+eRqT6Fo3x+/YXef9MGbjpSqX5OTzGKfftfu4.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ED25519) to the list of known hosts.
dion@localhost's password:
Linux kali 5.19.0-kali2-amd64 #1 SMP PREEMPT_DYNAMIC Debian 5.19.11-1kali2 (2022-10-10) x86_64

The programs included with the Kali GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*copyright.

Kali GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.

```

## 7. Close the session.

```

(dion㉿kali)-[~]
└─$ exit
logout
Connection to localhost closed.

```

8. Download Hadoop from the below link as shown in the image:

<https://archive.apache.org/dist/hadoop/common/hadoop-2.9.0/> 9. After downloading the file, extract the

tar file.

```
(dion㉿kali)-[~/home/kali/Downloads]
└─$ ll
total 358156
-rwxrwxrwx 1 kali kali 366744329 Dec  2 05:15 hadoop-2.9.0.tar.gz

(dion㉿kali)-[~/home/kali/Downloads]
└─$ sudo tar -xvzf hadoop-2.9.0.tar.gz
[sudo] password for dion:
hadoop-2.9.0/
hadoop-2.9.0/include/
hadoop-2.9.0/include/Pipes.hh
hadoop-2.9.0/include/SerialUtils.hh
hadoop-2.9.0/include/hdfs.h
hadoop-2.9.0/include/StringUtils.hh
hadoop-2.9.0/include/TemplateFactory.hh
hadoop-2.9.0/NOTICE.txt
hadoop-2.9.0/lib/
hadoop-2.9.0/lib/native/
hadoop-2.9.0/lib/native/libhdfs.a
hadoop-2.9.0/lib/native/libhadoop.so
```

10. Move all the content in the file to the “hadoop” file in the local system. As well as give rights for user “dion” to access it.

```

(dion㉿kali)-[~/home/kali/Downloads]
└─$ ll
total 358160
drwxrwxrwx 9 kali kali 4096 Nov 13 2017 hadoop-2.9.0
-rwxrwxrwx 1 kali kali 366744329 Dec 2 05:15 hadoop-2.9.0.tar.gz

(dion㉿kali)-[~/home/kali/Downloads]
└─$ sudo mv hadoop-2.9.0 /usr/local/hadoop

(dion㉿kali)-[~/home/kali/Downloads]
└─$ ll hadoop-2.9.0
ls: cannot access 'hadoop-2.9.0': No such file or directory

(dion㉿kali)-[~/home/kali/Downloads]
└─$ ll
total 358156
-rwxrwxrwx 1 kali kali 366744329 Dec 2 05:15 hadoop-2.9.0.tar.gz

(dion㉿kali)-[~/home/kali/Downloads]
└─$ sudo chown -R dion /usr/local

(dion㉿kali)-[~/home/kali/Downloads]
└─$ █

```

11. Add Configurations in the profile of the system (“bashrc”).

```

(dion㉿kali)-[~/home/kali/Downloads]
└─$ sudo vi ~/.bashrc

```

The Configurations are:

```

export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_341 export
JAVA_PATH=$PATH:$JAVA_HOME/bin export
HADOOP_HOME=/usr/local/hadoop export
PATH=$PATH:$HADOOP_HOME/bin export
PATH=$PATH:$HADOOP_HOME/sbin 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=$HADOOP_HOME/native export
HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/native"

```

Run “source ~/.bashrc” to enable the changes.

```

(dion㉿kali)-[~/home/kali/Downloads]
└─$ source ~/.bashrc

```

12. Add the Java Path in Hadoop environment file

```

(dion㉿kali)-[~/home/kali/Downloads]
└─$ sudo vi /usr/local/hadoop/etc/hadoop/hadoop-env.sh
# The java implementation to use.
#export JAVA_HOME=${JAVA_HOME}
export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_341

```

13. Make changes in “core-site.xml”.

#### **core-site.xml:**

The core-site.xml file contains information such as the port number used for Hadoop instance, memory allocated for file system, memory limit for storing data, and the size of Read/Write buffers.

Core-Site.xml configuration:

```
<property>
<name>fs.default.name</name>
<value>hdfs://localhost:9000</value>
</property>
```

```
(dion@kali)-[/usr/local/hadoop/etc/hadoop]
$ vi core-site.xml

<configuration>
  <property>
    <name>fs.default.name</name>
    <value>hdfs://localhost:9000</value>
  </property>
</configuration>
~
```

14. Make changes in “hdfs-site.xml”.

#### hdfs-site.xml:

The hdfs-site.xml file contains information such as the value of replication data, namenode path, and datanode path of your local file systems, where you want to store the Hadoop infrastructure.

#### HDFS-Site.xml configuration:

```
<property>
<name>dfs.replication</name>
<value>1</value>
</property>
<property>
<name>dfs.name.name.dir</name>
<value>file:/usr/local/hadoop_tmp/hdfs/namenode</value>
</property>
<property>
<name>dfs.data.data.dir</name>
<value>file:/usr/local/hadoop_tmp/hdfs/datanode</value>
</property>
```

```
(dion@kali)-[/usr/local/hadoop/etc/hadoop]
$ vi hdfs-site.xml

<configuration>
  <property>
    <name>dfs.replication</name>
    <value>1</value>
  </property>
  <property>
    <name>dfs.name.name.dir</name>
    <value>file:/usr/local/hadoop_tmp/hdfs/namenode</value>
  </property>
  <property>
    <name>dfs.data.data.dir</name>
    <value>file:/usr/local/hadoop_tmp/hdfs/datanode</value>
  </property>
</configuration>
```

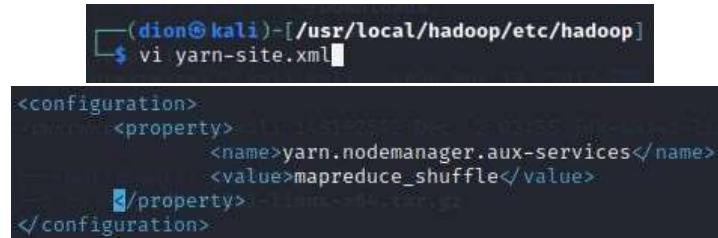
15. Make changes in “yarn-site.xml”.

### yarn-site.xml:

This file is used to configure yarn into Hadoop.

YARN-Site.xml configuration:

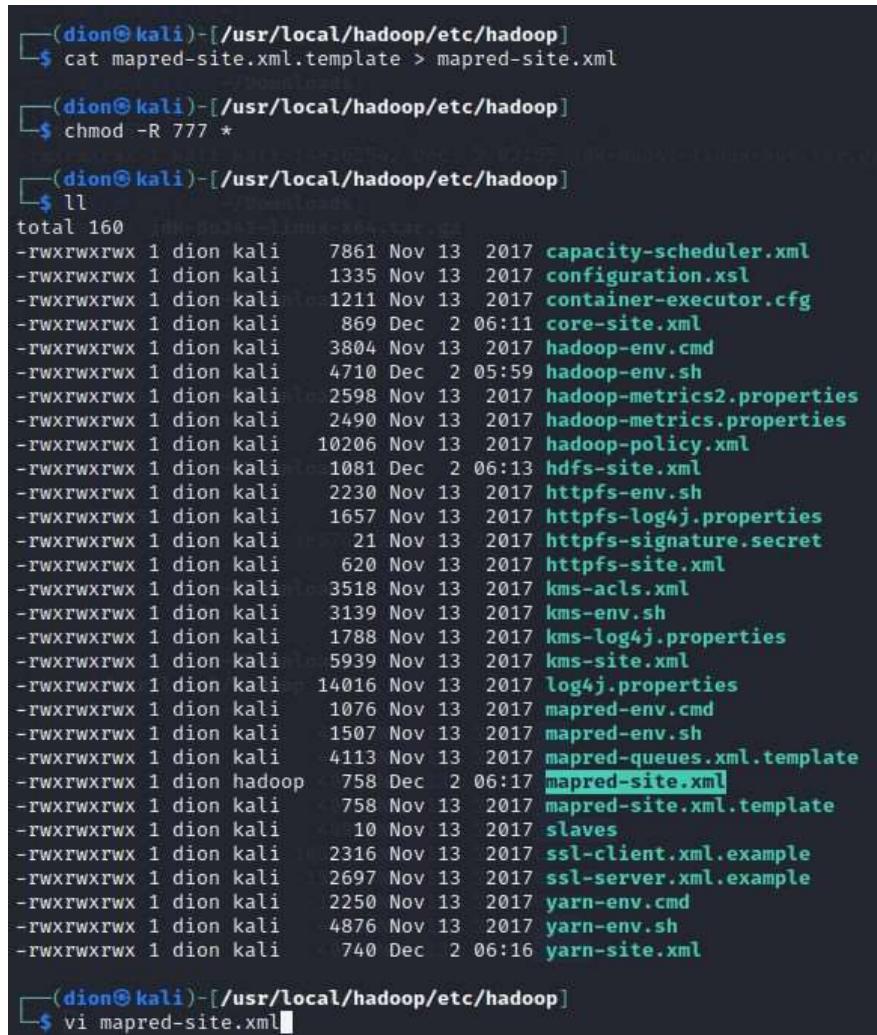
```
<property>
<name>yarn.nodemanager.aux-services</name>
<value>mapreduce_shuffle</value>
</property>
```



```
<configuration>
    <property>
        <name>yarn.nodemanager.aux-services</name>
        <value>mapreduce_shuffle</value>
    </property>
</configuration>
```

16. Now we must configure the mapred-site.xml.

In order to do this we will have to create the file “mapred-site.xml” by copy it from “mapred-site.xml.template” as show in the image below.



```
<(dion@kali)-[/usr/local/hadoop/etc/hadoop]>
$ cat mapred-site.xml.template > mapred-site.xml
<(dion@kali)-[/usr/local/hadoop/etc/hadoop]>
$ chmod -R 777 *
<(dion@kali)-[/usr/local/hadoop/etc/hadoop]>
$ ll
total 160
-rwxrwxrwx 1 dion kali    7861 Nov 13  2017 capacity-scheduler.xml
-rwxrwxrwx 1 dion kali    1335 Nov 13  2017 configuration.xsl
-rwxrwxrwx 1 dion kali   1211 Nov 13  2017 container-executor.cfg
-rwxrwxrwx 1 dion kali     869 Dec  2 06:11 core-site.xml
-rwxrwxrwx 1 dion kali   3804 Nov 13  2017 hadoop-env.cmd
-rwxrwxrwx 1 dion kali   4710 Dec  2 05:59 hadoop-env.sh
-rwxrwxrwx 1 dion kali   2598 Nov 13  2017 hadoop-metrics2.properties
-rwxrwxrwx 1 dion kali   2490 Nov 13  2017 hadoop-metrics.properties
-rwxrwxrwx 1 dion kali   10206 Nov 13  2017 hadoop-policy.xml
-rwxrwxrwx 1 dion kali   1081 Dec  2 06:13 hdfs-site.xml
-rwxrwxrwx 1 dion kali   2230 Nov 13  2017 httpfs-env.sh
-rwxrwxrwx 1 dion kali   1657 Nov 13  2017 httpfs-log4j.properties
-rwxrwxrwx 1 dion kali     21 Nov 13  2017 httpfs-signature.secret
-rwxrwxrwx 1 dion kali     620 Nov 13  2017 httpfs-site.xml
-rwxrwxrwx 1 dion kali   3518 Nov 13  2017 kms-acls.xml
-rwxrwxrwx 1 dion kali   3139 Nov 13  2017 kms-env.sh
-rwxrwxrwx 1 dion kali   1788 Nov 13  2017 kms-log4j.properties
-rwxrwxrwx 1 dion kali   5939 Nov 13  2017 kms-site.xml
-rwxrwxrwx 1 dion kali   14016 Nov 13  2017 log4j.properties
-rwxrwxrwx 1 dion kali    1076 Nov 13  2017 mapred-env.cmd
-rwxrwxrwx 1 dion kali   1507 Nov 13  2017 mapred-env.sh
-rwxrwxrwx 1 dion kali   4113 Nov 13  2017 mapred-queues.xml.template
-rwxrwxrwx 1 dion hadoop    758 Dec  2 06:17 mapred-site.xml
-rwxrwxrwx 1 dion kali    758 Nov 13  2017 mapred-site.xml.template
-rwxrwxrwx 1 dion kali     10 Nov 13  2017 slaves
-rwxrwxrwx 1 dion kali   2316 Nov 13  2017 ssl-client.xml.example
-rwxrwxrwx 1 dion kali   2697 Nov 13  2017 ssl-server.xml.example
-rwxrwxrwx 1 dion kali   2250 Nov 13  2017 yarn-env.cmd
-rwxrwxrwx 1 dion kali   4876 Nov 13  2017 yarn-env.sh
-rwxrwxrwx 1 dion kali    740 Dec  2 06:16 yarn-site.xml

<(dion@kali)-[/usr/local/hadoop/etc/hadoop]>
$ vi mapred-site.xml
```

```

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

```

```

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

```

17. Next, we will create the following directories and transfer the ownership to the “dion”. These directories are where hadoop will store data about the namenode and the datanode.

```

└──(dion㉿kali)-[~/]
  └──$ sudo mkdir -p /usr/local/hadoop_space
[sudo] password for dion:
Sorry, try again.
[sudo] password for dion:

└──(dion㉿kali)-[~/]
  └──$ sudo mkdir -p /usr/local/hadoop_space/hdfs/namenode

└──(dion㉿kali)-[~/]
  └──$ sudo mkdir -p /usr/local/hadoop_space/hdfs/datanode

└──(dion㉿kali)-[~/]
  └──$ sudo chown -R dion /usr/local/hadoop_space

└──(dion㉿kali)-[~/]
  └──$ █

```

18. Then we format the namenode in the HDFS file structure.

```

└──(dion㉿kali)-[~]
  └──$ hdfs namenode -format
22/12/02 06:26:36 INFO namenode.NameNode: STARTUP_MSG:
*****STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = kali/127.0.1.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 2.9.0
STARTUP_MSG: classpath = /usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/hadoop/common/lib/commons-digeste
4.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-math3-3.6.1.jar:/usr/local/hadoop/share/hadoop/commo
lf4j-api-1.7.25.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-lang3-3.6.jar:/usr/local/hadoop/share/hado
mon/lib/mockito-all-1.8.5.jar:/usr/local/hadoop/share/hadoop/common/lib/snappy-java-1.0.5.jar:/usr/local/hadoop/sha
re/hadoop/common/lib/yz-1.0.jar:/usr/local/hadoop/share/hadoop/common/lib/zookeeper-3.7.1.jar:/usr/local/hadoop/
/share/hadoop/common/lib/zookeeper-3.7.1.jar:/usr/local/hadoop/share/hadoop/zookeeper/zookeeper-3.7.1.jar

```

19. Next we will start the Distributed file system.

```

└──(dion㉿kali)-[~]
  └──$ start-dfs.sh
22/12/02 06:27:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built-in
n-java classes where applicable
Starting namenodes on [localhost]
dion@localhost's password:
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-dion-namenode-kali.out
dion@localhost's password:
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-dion-datanode-kali.out
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
ED25519 key fingerprint is SHA256:4d/a0Z+eRqT6F03x+/YXef9MgbJpSqX50TzGKfftFu4.
This host key is known by the following other names/addresses:
  ~/.ssh/known_hosts:1: [hashed name]
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
0.0.0.0: Warning: Permanently added '0.0.0.0' (ED25519) to the list of known hosts.
dion@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-dion-secondarynamenode-kali.out
22/12/02 06:31:35 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built-in
n-java classes where applicable

```

20. Followed by the YARN system.

```
(dion㉿kali)-[~]
└─$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-dion-resourcemanager-kali.out
dion@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-dion-nodemanager-kali.out
```

21. Run “start-all.sh” to run all daemons and Check if all the daemons are running

```
(dion㉿kali)-[~]
└─$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
22/12/02 06:48:32 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built-in java classes where applicable
Starting namenodes on [localhost]
dion@localhost's password:
localhost: namenode running as process 56030. Stop it first.
dion@localhost's password:
localhost: datanode running as process 56213. Stop it first.
Starting secondary namenodes [0.0.0.0]
dion@0.0.0.0's password:
0.0.0.0: secondarynamenode running as process 56714. Stop it first.
22/12/02 06:48:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built-in java classes where applicable
starting yarn daemons
resourcemanager running as process 57508. Stop it first.
dion@localhost's password:
localhost: nodemanager running as process 58225. Stop it first.

(dion㉿kali)-[~]
└─$ jps
61600 Jps
58225 NodeManager
57508 ResourceManager
56213 DataNode
56714 SecondaryNameNode
56030 NameNode
```

22. As we can see that Hadoop has been successfully installed and activated.

We must go the URL <http://localhost:500070> to access the Hadoop Server page.

The screenshot shows a web browser window titled "Namenode information". The address bar displays "localhost:50070/dfshealth.html#tab-overview". The page content is titled "Hadoop" and "Overview". It includes a table with cluster statistics:

Started:	Fri Dec 02 06:28:25 -0500 2022
Version:	2.9.0, r756ebc8394e473ac25feac05fa493f6d612e6c50
Compiled:	Mon Nov 13 18:15:00 -0500 2017 by arsuresh from branch-2.9.0
Cluster ID:	CID-8e57a896-3592-4781-b697-422b95612fb
Block Pool ID:	BP-1712706018-127.0.1.1-1669980401181

Below the table, there is a "Summary" section with the following text:

Security is off.  
Safemode is off.  
1 files and directories, 0 blocks = 1 total filesystem object(s).  
Heap Memory used 33.48 MB of 172 MB Heap Memory. Max Heap Memory is 889 MB.  
Non Heap Memory used 46.34 MB of 47.4 MB Committed Non Heap Memory. Max Non Heap Memory is <unbounded>.

We can also see the information of the cluster.

The screenshot shows the Hadoop NameNode information interface at [localhost:8088/cluster](http://localhost:8088/cluster). The left sidebar has a tree view with 'Cluster' expanded, showing 'About', 'Nodes', 'Node Labels', 'Applications' (with sub-options: NEW, NEW\_SAVING, SUBMITTED, ACCEPTED, RUNNING, FINISHED, FAILED, KILLED), 'Scheduler', and 'Tools'. The main area displays 'Cluster Metrics' with zero values for all metrics. It also shows 'Cluster Nodes Metrics' with one active node and zero decommissioning or decommissioned nodes. Under 'Scheduler Metrics', it shows 'Capacity Scheduler' with 'MEMORY' as the scheduling resource type and a minimum allocation of <memory:1024, vCores:1>. A table for 'Scheduler Metrics' is shown with columns: ID, User, Name, Application Type, Queue, Application Priority, StartTime, FinishTime, State, FinalStatus, and Run Con. The message 'No data available' is displayed below the table. At the bottom, it says 'Showing 0 to 0 of 0 entries'.

## Hadoop Version

```
(dion㉿kali)-[~]
└─$ cd /usr/local
(dion㉿kali)-[/usr/local]
└─$ hadoop version
Hadoop 2.9.0
Subversion https://git-wip-us.apache.org/repos/asf/hadoop.git -r 756ebc8394e473ac25feac05fa493f6d612e6c50
Compiled by arsuresh on 2017-11-13T23:15Z
Compiled with protoc 2.5.0
From source with checksum 0a76a9a32a5257331741f8d5932f183
This command was run using /usr/local/hadoop/share/hadoop/common/hadoop-common-2.9.0.jar
```

## Java Version

```
(dion㉿kali)-[/usr/local]
└─$ java -version
java version "1.8.0_341"
Java(TM) SE Runtime Environment (build 1.8.0_341-b10)
Java HotSpot(TM) 64-Bit Server VM (build 25.341-b10, mixed mode)

(dion㉿kali)-[/usr/local]
└─$ javac -version
javac 1.8.0_341
```

## C. HBASE:

1. Download HBASE from the given link below as shown in the image. (*Note: choose stable version and choose bin*) <https://hbase.apache.org/downloads.html>

Namenode information × All Applications × Apache HBase – Apache HBase +

Kali Linux Kali Tools Kali Docs Kali Forums Kali NetHunter Exploit-DB Google Hacking DB OffSec

Apache HBase Project Project Information Documentation and API ASF

published KEYS file. See Verify The Integrity Of The Files for how to verify your mirrored downloads.

## Releases

Version	Release Date	Compatibility Report	Changes	Release Notes	Download	Notices
3.0.0-alpha-3	2022/06/27	<a href="#">3.0.0-alpha-3 vs 2.0.0</a>	<a href="#">Changes</a>	<a href="#">Release Notes</a>	<a href="#">src (sha512 asc)</a> <a href="#">bin (sha512 asc)</a> <a href="#">client-bin (sha512 asc)</a>	Testing only, not production ready
2.5.1	2022/10/28	<a href="#">2.5.1 vs 2.5.0</a>	<a href="#">Changes</a>	<a href="#">Release Notes</a>	<a href="#">src (sha512 asc)</a> <a href="#">bin (sha512 asc)</a> <a href="#">client-bin (sha512 asc)</a>	
2.4.15	2022/10/28	<a href="#">2.4.15 vs 2.4.14</a>	<a href="#">Changes</a>	<a href="#">Release Notes</a>	<a href="#">src (sha512 asc)</a> <a href="#">bin (sha512 asc)</a> <a href="#">client-bin (sha512 asc)</a>	stable release

## Connectors

Version	Release Date	Compatibility Report	Changes	Release Notes	Download	Notices
1.0.0	2019/05/03		<a href="#">Changes</a>	<a href="#">Release Notes</a>	<a href="#">src (sha512 asc)</a> <a href="#">bin (sha512 asc)</a>	

<https://www.apache.org/dyn/closer.lua/hbase/2.4.15/hbase-2.4.15-bin.tar.gz>

```
(dion㉿kali)-[~/Downloads]
$ ll
total 277576
-rwxrwxrwx 1 kali kali 284230281 Dec  2 09:04 hbase-2.4.15-bin.tar.gz
```

2. Extract the files and move them to the “hbase” folder in local system while giving right for user “dion”.

```
(dion㉿kali)-[~/Downloads]
$ sudo tar -xvzf hbase-2.4.15-bin.tar.gz
hbase-2.4.15/LICENSE.txt
hbase-2.4.15/NOTICE.txt
hbase-2.4.15/LEGAL
hbase-2.4.15/docs/
hbase-2.4.15/docs/apidocs/
hbase-2.4.15/docs/apidocs/org/
hbase-2.4.15/docs/apidocs/org/apache/
hbase-2.4.15/docs/apidocs/org/apache/hadoop/
hbase-2.4.15/docs/apidocs/org/apache/hadoop/hbase/
hbase-2.4.15/docs/apidocs/org/apache/hadoop/hbase/backup/
hbase-2.4.15/docs/apidocs/org/apache/hadoop/hbase/chaos/
hbase-2.4.15/docs/apidocs/org/apache/hadoop/hbase/chaos/class-use/
hbase-2.4.15/docs/apidocs/org/apache/hadoop/hbase/class-use/
```

```
(dion㉿kali)-[~/Downloads]
$ ll
total 277580
drwxrwxrwx 7 root root        4096 Dec  2 09:24 hbase-2.4.15
-rwxrwxrwx 1 kali kali 284230281 Dec  2 09:04 hbase-2.4.15-bin.tar.gz
```

```
(dion㉿kali)-[~/Downloads]
$ sudo mv hbase-2.4.15 /usr/local/hbase
(dion㉿kali)-[~/Downloads]
$ sudo chown -R dion /usr/local
```

3. Add these Configuration in “bashrc” for HBASE/

```
(dion㉿kali)-[~/Downloads]
$ sudo vi ~/.bashrc
```

Configurations:

```
export HBASE_HOME=/usr/local/hbase
export PATH=$PATH:$HBASE_HOME/bin
```

```
export HADOOP_OPTS="-Djava.library.path"
export HBASE_HOME=/usr/local/hbase
export PATH=$PATH:$HBASE_HOME/bin
```

4. “source” to save and enable the changes.

```
(dion㉿kali)-[~/home/kali/Downloads]
$ source ~/.bashrc
```

5. Add Java path in HBASE environment file.

```
(dion㉿kali)-[/usr/local/hbase/conf]
$ vi hbase-env.sh

# The java implementation to use. Java 1.8+ required.
export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_341
```

6. Add these configurations in hbase-site.xml.

HBASE-site.xml:

```
<property>
  <name>hbase.rootdir</name>
  <value>file:/usr/local/hbase/hfiles</value>
</property>
<property>
  <name>hbase.zookeeper.property.dataDir</name>
  <value>/usr/local/zookeeper</value>
</property>
<property>
  <name>hbase.tmp.dir</name>
  <value>./tmp</value>
</property>
<property>
  <name>hbase.unsafe.stream.capability.enforce</name>
  <value>false</value>
</property>
<property>
  <name>hbase.cluster.distributed</name>
  <value>true</value>
</property>
<property>
  <name>hbase.rootdir</name>
  <value>hdfs://localhost:9000/hbase</value>
</property>
```

```
(dion㉿kali)-[/usr/local/hbase/conf]
└─$ vi hbase-site.xml

<property>
    <name>hbase.rootdir</name>
    <value>file:/usr/local/hbase/hfiles</value>
</property>
<property>
    <name>hbase.zookeeper.property.dataDir</name>
    <value>/usr/local/zookeeper</value>
</property>
<property>
    <name>hbase.tmp.dir</name>
    <value>./tmp</value>
</property>
<property>
    <name>hbase.unsafe.stream.capability.enforce</name>
    <value>false</value>
</property>
<property>
    <name>hbase.cluster.distributed</name>
    <value>true</value>
</property>
<property>
    <name>hbase.rootdir</name>
    <value>hdfs://localhost:9000/hbase</value>
</property>
</configuration>
```

## 7. Start HBASE.

```
(dion㉿kali)-[/usr/local/hbase/conf]
└─$ cd /usr/local/hbase/bin

(dion㉿kali)-[/usr/local/hbase/bin]
└─$ ./start-hbase.sh
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/client-facing-thirdparty/slf4j-reload4j-1.7.33.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.slf4j.impl.Log4jLoggerFactory]
running master, logging to /usr/local/hbase/logs/hbase-dion-master-kali.out
```

## 8. Check if all the daemons are running especially “HMaster”.

```
(dion㉿kali)-[/usr/local/hbase/conf]
└─$ jps
35026 HQuorumPeer
35142 HMaster
17942 DataNode
17718 NameNode
14631 ResourceManager
35271 HRegionServer
18554 NodeManager
35468 Jps
18141 SecondaryNameNode
```

## 9. To run HBASE shell, type “hbase shell”.

```
(dion㉿kali)-[~]
└─$ hbase shell
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/client-facing-thirdparty/slf4j-reload4j-1.7.33.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.slf4j.impl.Log4jLoggerFactory]
2022-12-04 23:04:02,937 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
HBase Shell
Use "help" to get list of supported commands.
Use "exit" to quit this interactive shell.
For Reference, please visit: http://hbase.apache.org/2.0/book.html#shell
Version 2.4.15, r35310fcdb611a1d04d75eb7db2e592dd34e4d5b6, Thu Oct 13 11:42:20 PDT 2022
Took 0.0032 seconds
hbase:001:0> version
2.4.15 - r35310fcdb611a1d04d75eb7db2e592dd34e4d5b6, Thu Oct 13 11:42:20 PDT 2022
Took 0.0004 seconds
hbase:002:0> status
1 active master, 0 backup masters, 1 servers, 0 dead, 2.0000 average load
Took 2.8305 seconds
hbase:003:0> 
```

## 10. Since HBASE is working. We can check the web interface by going to the

URL <http://localhost:16010/master-status>

The screenshot shows the Apache HBase master status interface. At the top, there are tabs for 'Namenode information', 'All Applications', 'Apache HBase - Apache H', and 'Backup Master: kali'. The main content area has a header 'Backup Master kali'. Below it, a message says 'Current Active Master: kali'. A 'Tasks' section contains a table with one row: 'Fri Dec 02 13:47:11 EST 2022' (Start Time), 'Master startup' (Description), 'RUNNING (since 49sec ago)' (State), and 'Initializing Master file system (since 41sec ago)' (Status). There are also buttons for 'Show All Monitored Tasks', 'Show non-RPC Tasks' (which is selected), 'Show All RPC Handler Tasks', 'Show Active RPC Calls', and 'Show Client Operations'. Below the tasks is a 'Software Attributes' section with a table. The columns are 'Attribute Name', 'Value', and 'Description'. There are no visible entries.

### Implementation:

Here we will create a “Student” table with details such as

- a. S\_ID,
- b. S\_NAME(where we will have F\_NAME and L\_NAME),
- c. DOB,
- d. GENDER,
- e. S\_ADDRESS,
- f. S\_PHONE\_NO

Table representation:

S_ID	S_NAME		DOB	GENDER	S_ADDRESS	S_PHONE_NO
	F_NAME	L_NAME				
CS21001	DION	FERNANDES	01-04-2000	M	BANDRA	9999990001
CS21002	RUDRA	RAO	05-12-1998	M	MAHIM	9999990002
CS21003	ZAID	SHAIKH	23-07-2000	M	SION	9999990003
CS21004	RAJ	KUMAR	30-11-1999	M	DADAR	9999990004

1. Run “version” and “status” to check if HBASE is running.

```

hbase:001:0> version
2.4.15, r35310fc6b11a1d04d75eb7db2e592dd34e4d5b6, Thu Oct 13 11:42:20 PDT 2022
Took 0.0004 seconds
hbase:002:0> status
1 active master, 0 backup masters, 1 servers, 0 dead, 2.0000 average load
Took 2.8305 seconds

```

2. Run “list” to list down the tables available.

```

hbase:004:0> list
TABLE
0 row(s)
Took 0.1644 seconds
⇒ []
hbase:005:0>

```

3. Create table “Student” and check if it is created.

```

hbase:006:0>
hbase:006:0> create 'Student','S_ID','S_NAME','DOB','GENDER','S_ADDRESS','S_PHONE_NO'
Created table Student
Took 1.4548 seconds
⇒ Hbase::Table - Student
hbase:007:0> list
TABLE
Student
1 row(s)
Took 0.0303 seconds
⇒ ["Student"]
hbase:008:0>

```

4. Run “describe ‘Student’”

```

hbase:010:0> describe 'Student'
Table Student is ENABLED
  ↪ S_ID, S_NAME, DION
  ↪ S_NAME, FERNANDES
COLUMN FAMILIES DESCRIPTION
{NAME => 'DOB', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', COMPRESSION => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
{NAME => 'GENDER', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', COMPRESSION => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
{NAME => 'S_ADDRESS', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', COMPRESSION => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
{NAME => 'S_ID', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', COMPRESSION => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
{NAME => 'S_NAME', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', COMPRESSION => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
{NAME => 'S_PHONE_NO', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', COMPRESSION => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
6 row(s)
Quota is disabled
Took 1.1976 seconds
hbase:011:0>

```

5. Now we will add data in the table ‘Student’. Code: put

```

'Student','CS21001','S_NAME:F_NAME','DION' put
'Student','CS21001','S_NAME:L_NAME','FERNANDES'
put 'Student','CS21001','DOB','01-04-2000' put
'Student','CS21001','GENDER','M' put
'Student','CS21001','S_ADDRESS','BANDRA'
put 'Student','CS21001','S_PHONE_NO','9999990001'

put 'Student','CS21002','S_NAME:F_NAME','RUDRA' put
'Student','CS21002','S_NAME:L_NAME','RAO' put
'Student','CS21002','DOB','05-12-1998' put
'Student','CS21002','GENDER','M' put
'Student','CS21002','S_ADDRESS','MAHIM'
put 'Student','CS21002','S_PHONE_NO','9999990002'

```

```

put 'Student','CS21003','S_NAME:F_NAME','ZAID' put
'Student','CS21003','S_NAME:L_NAME','SHAIKH'
put 'Student','CS21003','DOB','23-07-2000' put
'Student','CS21003','GENDER','M' put
'Student','CS21003','S_ADDRESS','SION'
put 'Student','CS21003','S_PHONE_NO','9999990003'

put 'Student','CS21004','S_NAME:F_NAME','RAJ' put
'Student','CS21004','S_NAME:L_NAME','KUMAR'
put 'Student','CS21004','DOB','30-11-1999' put
'Student','CS21004','GENDER','M' put
'Student','CS21004','S_ADDRESS','DADAR' put
'Student','CS21004','S_PHONE_NO','9999990004'

```

```

hbase:011:0> put 'Student','CS21001','S_NAME:F_NAME','DION'
Took 0.9969 seconds
hbase:012:0> put 'Student','CS21001','S_NAME:L_NAME','FERNANDES'
Took 0.0284 seconds
hbase:013:0> put 'Student','CS21001','DOB','01-04-2000'
Took 0.0131 seconds
hbase:014:0> put 'Student','CS21001','GENDER','M'
Took 0.1835 seconds
hbase:015:0> put 'Student','CS21001','S_ADDRESS','BANDRA'
Took 0.1031 seconds
hbase:016:0> put 'Student','CS21001','S_PHONE_NO','9999990001'
Took 0.0129 seconds
hbase:017:0> scan 'Student'
ROW                                     COLUMN+CELL
CS21001                                column=DOB:, timestamp=2022-12-04T23:31:44.651, value=01-04-2000
CS21001                                column=GENDER:, timestamp=2022-12-04T23:31:44.939, value=M
CS21001                                column=S_ADDRESS:, timestamp=2022-12-04T23:31:45.247, value=BANDRA
CS21001                                column=S_NAME:F_NAME, timestamp=2022-12-04T23:31:19.390, value=DION
CS21001                                column=S_NAME:L_NAME, timestamp=2022-12-04T23:31:44.412, value=FERNANDES
CS21001                                column=S_PHONE_NO:, timestamp=2022-12-04T23:31:49.310, value=9999990001
1 row(s)
Took 0.2246 seconds
hbase:018:0>

```

## 6. Run “scan ‘Student’” to display the table values.

```

hbase:048:0> scan 'Student'
ROW                                     COLUMN+CELL
CS21001                                column=DOB:, timestamp=2022-12-04T23:31:44.651, value=01-04-2000
CS21001                                column=GENDER:, timestamp=2022-12-04T23:31:44.939, value=M
CS21001                                column=S_ADDRESS:, timestamp=2022-12-04T23:31:45.247, value=BANDRA
CS21001                                column=S_NAME:F_NAME, timestamp=2022-12-04T23:31:19.390, value=DION
CS21001                                column=S_NAME:L_NAME, timestamp=2022-12-04T23:31:44.412, value=FERNANDES
CS21001                                column=S_PHONE_NO:, timestamp=2022-12-04T23:31:49.310, value=9999990001
CS21002                                column=DOB:, timestamp=2022-12-04T23:33:25.567, value=05-12-1998
CS21002                                column=GENDER:, timestamp=2022-12-04T23:33:26.006, value=M
CS21002                                column=S_ADDRESS:, timestamp=2022-12-04T23:33:26.425, value=MAHIM
CS21002                                column=S_NAME:F_NAME, timestamp=2022-12-04T23:33:23.722, value=RUDRA
CS21002                                column=S_NAME:L_NAME, timestamp=2022-12-04T23:33:25.067, value=RAO
CS21002                                column=S_PHONE_NO:, timestamp=2022-12-04T23:33:26.561, value=9999990002
CS21003                                column=DOB:, timestamp=2022-12-04T23:33:27.034, value=23-07-2000
CS21003                                column=GENDER:, timestamp=2022-12-04T23:33:27.201, value=M
CS21003                                column=S_ADDRESS:, timestamp=2022-12-04T23:33:27.262, value=SION
CS21003                                column=S_NAME:F_NAME, timestamp=2022-12-04T23:33:26.739, value=ZAID
CS21003                                column=S_NAME:L_NAME, timestamp=2022-12-04T23:33:26.849, value=SHAIKH
CS21003                                column=S_PHONE_NO:, timestamp=2022-12-04T23:33:27.357, value=9999990003
CS21004                                column=DOB:, timestamp=2022-12-04T23:33:27.614, value=30-11-1999
CS21004                                column=GENDER:, timestamp=2022-12-04T23:33:27.681, value=M
CS21004                                column=S_ADDRESS:, timestamp=2022-12-04T23:33:27.734, value=DADAR
CS21004                                column=S_NAME:F_NAME, timestamp=2022-12-04T23:33:27.487, value=RAJ
CS21004                                column=S_NAME:L_NAME, timestamp=2022-12-04T23:33:27.553, value=KUMAR
4 row(s)
Took 0.3028 seconds
hbase:049:0>

```

## 7. We can use “get” to get the values of specific row.

```

hbase:049:0> get 'Student','CS21001'
COLUMN                                     CELL
  DOB:                                         timestamp=2022-12-04T23:31:44.651, value=01-04-2000
  GENDER:                                       timestamp=2022-12-04T23:31:44.939, value=M
  S_ADDRESS:                                     timestamp=2022-12-04T23:31:45.247, value=BANDRA
  S_NAME:F_NAME                                timestamp=2022-12-04T23:31:19.390, value=DION
  S_NAME:L_NAME                                 timestamp=2022-12-04T23:31:44.412, value=FERNANDES
  S_PHONE_NO:                                    timestamp=2022-12-04T23:31:49.310, value=9999990001
1 row(s)
Took 0.3466 seconds
hbase:050:0> ■

Took 0.0001 seconds
hbase:003:0> get 'Student','CS21003','S_NAME:L_NAME'
COLUMN                                     CELL
  S_NAME:L_NAME                               timestamp=2022-12-04T23:33:26.849, value=SHAIKH
1 row(s)
Took 0.1788 seconds
hbase:004:0> ■

Took 0.0001 seconds
hbase:007:0> get 'Student','CS21004','S_PHONE_NO'
COLUMN                                     CELL
  S_PHONE_NO:                                timestamp=2022-12-04T23:53:38.742, value=9999990004\x0A
1 row(s)
Took 0.0312 seconds
hbase:008:0> ■

```

As seen in the image on top there is an invalid phone number value for Student “CS21004”. In order to change the value, we will update the value.

#### 8. To update the value use “put”.

```

hbase:008:0> put 'Student','CS21004','S_PHONE_NO','9999990444'
Took 0.0721 seconds
hbase:009:0> get 'Student','CS21004','S_PHONE_NO'
COLUMN                                     CELL
  S_PHONE_NO:                                timestamp=2022-12-04T23:58:34.702, value=9999990444
1 row(s)
Took 0.0613 seconds
hbase:010:0> ■

```

#### 9. Now we will perform “delete” operations.

##### a. Deleting the last name of Student “CS21002”. Code: delete

'Student','CS21002','S\_NAME:L\_NAME'

```

Took 0.2053 seconds
hbase:003:0> delete 'Student','CS21002','S_NAME:L_NAME'
Took 0.6181 seconds
hbase:004:0> get 'Student','CS21002','S_NAME:L_NAME'
COLUMN                                     CELL
  0 row(s)
Took 0.1921 seconds
hbase:005:0> get 'Student','CS21002'
COLUMN                                     CELL
  DOB:                                         timestamp=2022-12-04T23:33:25.567, value=05-12-1998
  GENDER:                                       timestamp=2022-12-04T23:33:26.006, value=M
  S_ADDRESS:                                     timestamp=2022-12-04T23:33:26.425, value=MAHIM
  S_NAME:F_NAME                                timestamp=2022-12-04T23:33:23.722, value=RUDRA
  S_PHONE_NO:                                    timestamp=2022-12-04T23:33:26.561, value=9999990002
1 row(s)
Took 0.1696 seconds
hbase:006:0> ■

```

##### b. Deleting the address of Student “CS21004”. Code: delete

'Student','CS21004','S\_ADDRESS'

```
hbase:008:0> delete 'Student','CS21004','S_ADDRESS:'  
Took 0.0292 seconds  
hbase:009:0> get 'Student','CS21004'  
COLUMN CELL  
DOB: timestamp=2022-12-04T23:33:27.614, value=30-11-1999  
GENDER: timestamp=2022-12-04T23:33:27.681, value=M  
S_NAME:F_NAME timestamp=2022-12-04T23:33:27.487, value=RAJ  
S_NAME:L_NAME timestamp=2022-12-04T23:33:27.553, value=KUMAR  
S_PHONE_NO: timestamp=2022-12-04T23:58:34.702, value=9999990444  
1 row(s)  
Took 0.0324 seconds  
hbase:010:0> █
```

c. Deleting the entire record of Student “CS21004”.

Code: deleteall 'Student','CS21004'

```
hbase:010:0> deleteall 'Student','CS21004'  
Took 0.1360 seconds  
hbase:011:0> get 'Student','CS21004' CELL  
COLUMN  
0 row(s)  
Took 0.0071 seconds  
hbase:012:0> █
```

10. Run “count” to count the number of rows in a table.

```
hbase:012:0> count 'Student'  
3 row(s)  
Took 0.3053 seconds  
⇒ 3
```

11. Truncating the table ‘Student’ will remove all the data from the table.

```
hbase:015:0> truncate 'Student'  
Truncating 'Student' table (it may take a while):  
Disabling table...  
Truncating table...  
Took 57.6435 seconds  
hbase:016:0> scan 'Student'  
ROW COLUMN+CELL  
0 row(s)  
Took 5.2316 seconds  
hbase:017:0> █
```

12. Dropping the table ‘Student’.

(Note: in order to delete the table, we will have to first disable it)

First, we check if the table is disabled or not.

```
hbase:017:0> is_disabled 'Student'  
false  
Took 0.0461 seconds  
⇒ false  
hbase:018:0> █
```

Since it is not disabled, we will have to disable it.

As seen in the image below, the table is now disabled.

```

    hbase:018:0> disable 'Student'
Took 1.4735 seconds
hbase:019:0> scan 'Student' S_PHONE_NO='9999999999'
ROW                                         COLUMN+CELL
org.apache.hadoop.hbase.TableNotEnabledException: Student is disabled.
    at org.apache.hadoop.hbase.client.ScannerCallable.prepare(ScannerCallable.java:157)
    at org.apache.hadoop.hbase.client.ScannerCallableWithReplicas$RetryingRPC.prepare(ScannerCallableWithReplicas.java:405)
    at org.apache.hadoop.hbase.client.RpcRetryingCallerImpl.callWithRetries(RpcRetryingCallerImpl.java:102)
    at org.apache.hadoop.hbase.client.ResultBoundedCompletionService$QueueingFuture.run(ResultBoundedCompletionService.java:74)
    at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1149)
    at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
    at java.lang.Thread.run(Thread.java:750)

ERROR: Table Student is disabled!
For usage try 'help "scan"'


Took 0.1341 seconds
hbase:020:0> ■

```

Since it is disabled, we can drop the table “Student”.

We can also use ‘exist’ to check if the table is deleted or no.

```

Took 0.1341 seconds
hbase:020:0> drop 'Student'
Took 1.4421 seconds
hbase:021:0> exists 'Student'
Table Student does not exist
Took 0.0875 seconds
⇒ false
hbase:022:0> ■

```

We can also use ‘list’ to check if the table is deleted or no.

```

hbase:022:0> list
TABLE
0 row(s) found
Took 0.3576 seconds
⇒ []
hbase:023:0> ■

```

## Conclusion:

In this project, we understood the basic requirements to Implement HBase. Such as:

1. Java:
  - a. Installation
  - b. Required Version for Hadoop
  - c. Environmental Configuration
2. Hadoop:
  - a. Installation
  - b. Environmental Configuration
  - c. Required Daemons
  - d. Displaying the Hadoop Server page as well as the Cluster page.
3. HBase:
  - a. Installation
  - b. Environmental Configuration
  - c. Implementation

In the Implementation, we have understood the requirements for HBase shell. As well as created a table “Student” while adding and manipulating the data init. We also understood few of the key words that can be used. Including few important notes to keep in mind.

## Practical 5

**Aim:** Perform importing and exporting of data between SQL and Hadoop using Sqoop.

### Sqoop Introductions

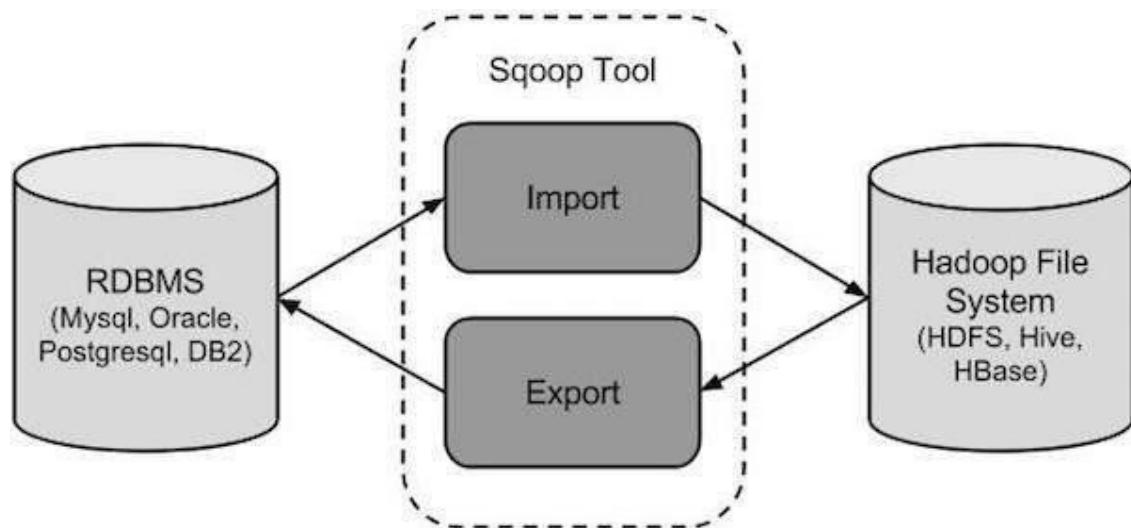
The traditional application management system, that is, the interaction of applications with relational database using RDBMS, is one of the sources that generate Big Data. Such Big Data, generated by RDBMS, is stored in Relational Database Servers in the relational database structure.

When Big Data storages and analyzers such as MapReduce, Hive, HBase, Cassandra, Pig, etc. of the Hadoop ecosystem came into picture, they required a tool to interact with the relational database servers for importing and exporting the Big Data residing in them. Here, Sqoop occupies a place in the Hadoop ecosystem to provide feasible interaction between relational database server and Hadoop's HDFS.

Sqoop – “SQL to Hadoop and Hadoop to SQL”

Sqoop is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases. It is provided by the Apache Software Foundation.

### Method of Sqoop



### Sqoop Import

The import tool imports individual tables from RDBMS to HDFS. Each row in a table is treated as a record in HDFS. All records are stored as text data in text files

### Sqoop Export

The export tool exports a set of files from HDFS back to an RDBMS. The files given as input to Sqoop contain records, which are called as rows in table. Those are read and parsed into a set of records and delimited with user-specified delimiter.

## Prerequisite for Sqoop Implementations

1. Installations of Java JDK.
2. Installations of Hadoop.
3. Installations of Sqoop package.

### 1. Installations of Java JDK

Step 1 - Update Kali

**Command:** - sudo apt update

```
(kali㉿kali)-[~] The JDK Software and Set JAVA_HOME
└─$ sudo apt update
[sudo] password for kali:
Get:1 http://ftp.harukasan.org/kali kali-rolling InRelease [30.5 kB]
Get:2 http://ftp.harukasan.org/kali kali-rolling/main amd64 Packages [17.7 MB]
Get:3 http://ftp.harukasan.org/kali kali-rolling/contrib amd64 Packages [108 kB]
Get:4 http://ftp.harukasan.org/kali kali-rolling/non-free amd64 Packages [199 kB]
Fetched 18.1 MB in 6min 39s (45.3 kB/s)
Reading package lists ... Done
Building dependency tree ... Done
Reading state information ... Done
308 packages can be upgraded. Run 'apt list --upgradable' to see them.
```

Step 2 - Install default JDK on Kali

**Command:** - sudo apt install default-jdk

```
(kali㉿kali)-[~] The JDK Software and Set JAVA_HOME on a UNIX System
└─$ sudo apt install default-jdk
[sudo] password for kali:
Reading package lists ... Done
Building dependency tree ... Done
Reading state information ... Done
The following packages were automatically installed and are no longer required:
  libexo-1-0 libpython3.8-dev libsane node-jquery python3.8-dev qt5-gtk2-platformtheme xfce4-mailwatch-plugin xfce4-s
  xfce4-weather-plugin
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
  default-jdk-headless libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libxau-dev libxcb1-dev libxdmcp-dev libx
  xproto-dev xorg-sgml-doctools xtrans-dev
Suggested packages:
  libice-doc libsm-doc libx11-doc libxcb-doc libxt-doc openjdk-11-demo openjdk-11-source visualvm
The following NEW packages will be installed:
  default-jdk default-jdk-headless libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libxau-dev libxcb1-dev libx
  xproto-dev xorg-sgml-doctools xtrans-dev
0 upgraded, 15 newly installed, 0 to remove and 308 not upgraded.
Need to get 226 MB of archives.
After this operation, 242 MB of additional disk space will be used.
Do you want to continue? [Y/n] y
Get:1 http://ftp.harukasan.org/kali kali-rolling/main amd64 openjdk-11-jdk-headless arm64 11.0-11+1 [217 MB]
```

Step 3 - Check the installation

**Command:** - java -version

```
(hadoopusr㉿kali)-[~]
$ java -version
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
openjdk version "11.0.17" 2022-10-18
OpenJDK Runtime Environment (build 11.0.17+8-post-Debian-2)
OpenJDK 64-Bit Server VM (build 11.0.17+8-post-Debian-2, mixed mode, sharing)
```

Step 4 – To update java jdk path in etc/environment path from root terminal.

**Command:** - sudo nano /etc/environment

```
(root㉿kali)-[~]
# sudo nano /etc/environment

GNU nano 6.4                               /etc/environment
# START KALI-DEFAULTS CONFIG
# Everything from here and until STOP KALI-DEFAULTS CONFIG
# was installed by the kali-defaults package, and it will
# be removed if ever the kali-defaults package is removed.
# If you want to disable a line, please do NOT remove it,
# as it would be added back when kali-defaults is upgraded.
# Instead, comment the line out, and your change will be
# preserved across upgrades.
PATH=/usr/local/sbin:/usr/local/bin:/usr/sbin:/usr/bin:/sbin:/bin:/usr/local/games:/usr/games
COMMAND_NOT_FOUND_INSTALL_PROMPT=1
POWERSHELL_UPDATECHECK=Off
POWERSHELL_TELEMETRY_OPTOUT=1
DOTNET_CLI_TELEMETRY_OPTOUT=1
# STOP KALI-DEFAULTS CONFIG

JAVA_HOME="/usr/lib/jvm/java-11-openjdk-amd64"

[ Read 16 lines ]
^G Help      ^O Write Out   ^W Where Is   ^K Cut        ^T Execute   ^C Location   M-U Undo
^X Exit      ^R Read File   ^N Replace    ^U Paste      ^J Justify    ^Y Go To Line  M-E Redo
```

Use Source to sink the file of etc/environment.

**Command:** - Source /etc/environment

```
(root㉿kali)-[~]
# source /etc/environment
```

## 2. Installations of Hadoop.

Step 1: - installing ssh server in root terminal.

**Command:** - apt-get install ssh

```
[root@kali]# apt-get install ssh
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following packages were automatically installed and are no longer required:
  freeglut3 libexporter-tiny-perl libhttp-server-simple-perl liblist-moreutils-perl
  liblist-moreutils-xs-perl libpython3.9-minimal libpython3.9-stdlib libwacom-bin python3-dataclasses-json
  python3-limiter python3-marshmallow-enum python3-mypy-extensions python3-responses python3-spyse
  python3-token-bucket python3-typing-inspect python3.9 python3.9-minimal
Use 'apt autoremove' to remove them.
The following NEW packages will be installed:
  ssh
0 upgraded, 1 newly installed, 0 to remove and 191 not upgraded.
Need to get 260 kB of archives.
After this operation, 276 kB of additional disk space will be used.
Get:1 http://kali.download/kali kali-rolling/main amd64 ssh all 1:9.0p1-1 [260 kB]
Fetched 260 kB in 1s (265 kB/s)
Selecting previously unselected package ssh.
(Reading database ... 338686 files and directories currently installed.)
Preparing to unpack .../ssh_1%3a9.0p1-1_all.deb ...
Unpacking ssh (1:9.0p1-1) ...
Setting up ssh (1:9.0p1-1) ...
```

Step 2: - Now, generate Public/Private rsa key pair:

**Command:** - ssh-keygen -t rsa -P ''

```
[root@kali]# ssh-keygen -t rsa -P ''
Generating public/private rsa key pair.
Enter file in which to save the key (/root/.ssh/id_rsa):
/root/.ssh/id_rsa already exists.
Overwrite (y/n)? y
Your identification has been saved in /root/.ssh/id_rsa
Your public key has been saved in /root/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:C7d8Bao+HHx3AnKkeBCpDdleqaN62q0YTc/S5NFprEI root@kali
The key's randomart image is:
+---[RSA 3072]---+
|   o . .
|   o + o .
|   = = o .
|   . *o+.o ..
|   E.o**oS. .
|   +.* +o=.oo ..
|   ..+ *..o+..o
|   .+.+ .o .
|ooo.. ...
+---[SHA256]---+
```

**Command:** - cat \$HOME/.ssh/id\_rsa.pub >> \$HOME/.ssh/authorized\_keys

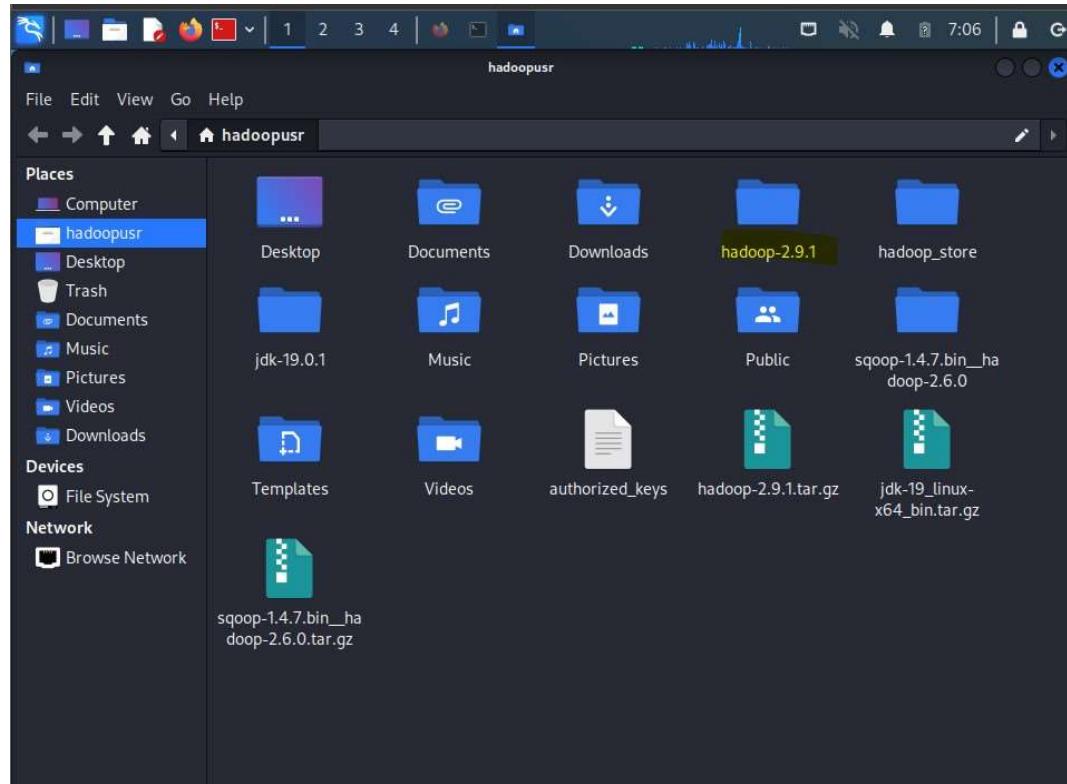
```
[root@kali]# cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
```

**Command:** - ssh localhost

Step 3: - Download Apache Hadoop 2.9.1 tar from below link

<https://hadoop.apache.org/release/2.9.1.html>

Once downloaded, Unzip in Home file.



Step 4: - Update Java JDK, Hadoop and other necessary path in .bashrc file.

**Command:** - Sudo nano .bashrc

```
(hadoopusr㉿kali)-[~]
$ sudo nano .bashrc
[sudo] password for hadoopusr:
```

```

export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export HADOOP_INSTALL=/home/hadoopusr/hadoop-2.9.1
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib"
#HADOOP VARIABLES END

```

**Command:** - source .bashrc

```

(hadoopusr㉿kali)-[~]
$ source .bashrc
(hadoopusr㉿kali)-[~]
$ 

```

Step 5: - Now, we need to configure some of Hadoop file below are the path and name.

Filepath:- /home/hadoopusr/hadoop-2.9.1/etc/hadoop/ hadoop-env.sh

```

~/hadoop-2.9.1/etc/hadoop/hadoop-env.sh - Mousepad
File Edit Search View Document Help
File Edit Search View Document Help
98 # Specify the JVM options to be used when starting the KBF routers.
99 # These options will be appended to the options specified as HADOOP_OPTS
100 # and therefore may override any similar flags set in HADOOP_OPTS
101 #
102 # export HADOOP_DFSROUTER_OPTS=""
103 ###
104 |
105 ###
106 # Advanced Users Only!
107 ###
108
109 # The directory where pid files are stored. /tmp by default.
110 # NOTE: this should be set to a directory that can only be written to by
111 #       the user that will run the hadoop daemons. Otherwise there is the
112 #       potential for a symlink attack.
113 export HADOOP_PID_DIR=${HADOOP_PID_DIR}
114 export HADOOP_SECURE_DN_PID_DIR=${HADOOP_PID_DIR}
115
116 # A string representing this instance of hadoop. $USER by default.
117 export HADOOP_IDENT_STRING=$USER
118
119 export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
120

```

We have to make few directories.

hadoop\_store – in same path where Hadoop-2.9.1/unzip Hadoop is there.

**Command:** - mkdir Hadoop\_store

**Command:** - cd hadoop\_store

Hdfs –under Hadoop\_store

**Command:** - mkdir hdfs

**Command:** - cd hdfs

Namenode – under hdfs

**Command:** - mkdir namenode

Datanode: - under hdfs

**Command:** - mkdir datanode

**File Path :-** /home/hadoopusr/hadoop-2.9.1/etc/hadoop/ hdfs-

**site.xml File Name:** - hdfs-site.xml place below code between configuration tag.

```
<property>
```

```
  <name>dfs.replication</name>
```

```
  <value>1</value>
```

```
  <description>Default block replication.
```

The actual number of replications can be specified when the file is created.

The default is used if replication is not specified in create time.

```
  </description>
```

```
</property>
```

```
<property>
```

```
  <name>dfs.namenode.name.dir</name>
```

```
  <value>file:#####NAMENODE_FOLDER_PATH#####</value>
```

```
</property>
```

```
<property>
```

```
  <name>dfs.datanode.data.dir</name>
```

```
  <value>file:#####DATANODE_FOLDER_PATH#####</value>
```

```
</property>
```

**File Path:** -/home/hadoopusr/hadoop-2.9.1/etc/hadoop/

**File Name:** - core-site.xml (For this, we have to create a Temp file in file path - /home/hadoopusr/hadoop-2.9.1/)

```
<property>
```

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

```

<value>#####TMP_FOLDER_PATH#####</value>
<description>A base for other temporary directories.</description>
</property>

<property>
<name>fs.default.name</name>
<value>hdfs://localhost:54310</value>
<description>The name of the default file system. A URI whose
scheme and authority determine the FileSystem implementation. The
uri's scheme determines the config property (fs.SCHEME.impl)
naming the FileSystem implementation class. The uri's authority is
used to determine the host, port, etc. for a filesystem.</description>
</property>

```

**File Path:** - /home/hadoopusr/hadoop-2.9.1/etc/hadoop/

**File Name:** - mapred-site.xml (This file won't be there, we need to copy mapredsite.xml.template ,paste it and rename it with mapred-site.xml)

for mapred-site.xml:

```

<property>
<name>mapred.job.tracker</name>
<value>localhost:54311</value>
<description>The host and port that the MapReduce job tracker
runs at. If "local", then jobs are run in-process as a single map
and reduce task. </description>
</property>

```

Step 6 :- Once above steps, are done. We need to run below command under Hadoop2.9.1 directory.

**Command:** - Hadoop namenode -format

Step 7: - once above commands is executed successfully, we need to start all daemons under bin directory

**Command:** - start-all.sh

```
(hadoopusr@kali)-[~]
$ cd hadoop-2.9.1

(hadoopusr@kali)-[~/hadoop-2.9.1]
$ cd bin

(hadoopusr@kali)-[~/hadoop-2.9.1/bin]
$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoop/usr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/04 07:49:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built-in java classes where applicable
Starting namenodes on [localhost]
localhost: namenode running as process 21023. Stop it first.
localhost: datanode running as process 21175. Stop it first.
Starting secondary namenodes [0.0.0.0]
0.0.0.0: starting secondarynamenode, logging to /home/hadoopusr/hadoop-2.9.1/logs/hadoop-hadoopusr-secondarynamenode-kali.out
0.0.0.0: WARNING: An illegal reflective access operation has occurred
0.0.0.0: WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
0.0.0.0: WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
0.0.0.0: WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
0.0.0.0: WARNING: All illegal access operations will be denied in a future release
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoop/usr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil

cancel()
0.0.0.0: WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
0.0.0.0: WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
0.0.0.0: WARNING: All illegal access operations will be denied in a future release
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoop/usr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/04 07:49:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built-in java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /home/hadoopusr/hadoop-2.9.1/logs/yarn-hadoopusr-resourcemanager-kali.out
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
localhost: starting nodemanager, logging to /home/hadoopusr/hadoop-2.9.1/logs/yarn-hadoopusr-nodemanager-kali.out
```

Step 8: - To stop all daemons.

**Command:** - stop-all.sh

```
(hadoopusr@kali:[~/hadoop-2.9.1/bin]
$ stop-all.sh
This script is Deprecated. Instead use stop-dfs.sh and stop-yarn.sh
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/04 07:56:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using built
in-java classes where applicable
Stopping namenodes on [localhost]
localhost: stopping namenode
localhost: stopping datanode
Stopping secondary namenodes [0.0.0.0]
0.0.0.0: stopping secondarynamenode
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/04 07:57:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform ... using built
in-java classes where applicable
stopping yarn daemons
stopping resourcemanager
localhost: stopping nodemanager
localhost: nodemanager did not stop gracefully after 5 seconds: killing with kill -9
no proxyserver to stop
```

### 3. Installations of Sqoop package.

Step 1:- First, we need to download scoop from below link.

<https://archive.apache.org/dist/sqoop/1.4.7/>

Step 2: - Unzip the sqoop tar file in home directories.

Step 3: - Need to download mysql-connector-java.jar and copy under below path file path.

/home/hadoopusr/sqoop-1.4.7-bin\_\_hadoop-2.6.0/Lib

Before starting to implement SQOOB we need to start all 5 daemons of Hadoop.

1. First we will start MySql Server
2. SQL server from below command in terminal.

```
(hadoopusr@kali)-[~]
$ service mysqld start
```

- Now we have to start MySql

```
(hadoopusr@kali)-[~]
$ sudo mysql -u root -p
Enter password:
Welcome to the MariaDB monitor. Commands end with ; or \g.
Your MariaDB connection id is 54
Server version: 10.6.10-MariaDB-1+b1 Debian n/a

Copyright (c) 2000, 2018, Oracle, MariaDB Corporation Ab and others.

Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.
```

- We will create a product table and insert few data so later we can export that table while implementing SQOOB.

```
MariaDB [tsest2]> create table Product( P_id int, P_Name varchar(20), P_Prices int(10), P_Discount Varchar(10));
Query OK, 0 rows affected (0.065 sec)

MariaDB [tsest2]> Insert into Product values(1001,"Table",10000,1000);
Query OK, 1 row affected (0.011 sec)

MariaDB [tsest2]> Insert into Product values(1002,"Chair",2000,200);
Query OK, 1 row affected (0.003 sec)

MariaDB [tsest2]> Insert into Product values(1003,"Laptop",20000,2000);
Query OK, 1 row affected (0.008 sec)

MariaDB [tsest2]> Insert into Product values(1004,"Mobile",7000,700);
Query OK, 1 row affected (0.003 sec)

MariaDB [tsest2]> Insert into Product values(1005,"TV",20000,2000);
Query OK, 1 row affected (0.001 sec)

MariaDB [tsest2]> select * from product;
ERROR 1146 (42S02): Table 'tsest2.product' doesn't exist
MariaDB [tsest2]> select * from Product;
+-----+-----+-----+-----+
| P_id | P_Name | P_Prices | P_Discount |
+-----+-----+-----+-----+
| 1001 | Table  | 10000   | 1000      |
| 1002 | Chair   | 2000    | 200       |
| 1003 | Laptop  | 20000   | 2000      |
| 1004 | Mobile  | 7000    | 700       |
| 1005 | TV      | 20000   | 2000      |
+-----+-----+-----+-----+
5 rows in set (0.001 sec)
```

- command to import data from MySql to HDFS(Hadoop).

sqoop import <----- tool

--connect jdbc:mysql://Localhost /Database name

--username root

-Password

--table table name **Example**

sqoop import --connect jdbc:mysql://localhost/tsest2 -username root -password  
kali --table Product -m1

```
(hadoopusr@kali)-[~/sqoop-1.4.7.bin_hadoop-2.6.0]
└─$ bin/sqoop import --connect jdbc:mysql://localhost/tsest2 --username root -password kali --table Product -m1
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../hcatalog does not exist! HCatalog jobs will fail.
Please set $CAT_HOME to the root of your HCatalog installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
22/12/03 10:12:17 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
22/12/03 10:12:17 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/12/03 10:12:18 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/12/03 10:12:18 INFO tool.CodeGenTool: Beginning code generation
Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.
22/12/03 10:12:19 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Product` AS t LIMIT 1
22/12/03 10:12:19 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Product` AS t LIMIT 1
22/12/03 10:12:19 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/hadoopusr/hadoop-2.9.1
Note: /tmp/sqoop-hadoopusr/compile/7f351cd81b6c24f0204a5dbc40bdce9/Product.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/12/03 10:12:25 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoopusr/compile/7f351cd81b6c24f0204a5dbc40bdce9/Product.jar
22/12/03 10:12:25 WARN manager.SqlManager: It looks like you are importing from mysql.
22/12/03 10:12:25 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/12/03 10:12:25 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/12/03 10:12:25 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/12/03 10:12:25 INFO mapreduce.ImportJobBase: Beginning import of Product
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 10:16:25 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 3 items
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 06:54 Persons
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 10:12 Product
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 07:28 Student
```

```
22/12/03 10:12:32 INFO mapred.Task: Task attempt_local2044251549_0001_m_000000_0 is allowed to commit now
22/12/03 10:12:32 INFO output.FileOutputCommitter: Saved output of task 'attempt_local2044251549_0001_m_000000_0'
Product/_temporary/_task_local2044251549_0001_m_000000
22/12/03 10:12:32 INFO mapred.LocalJobRunner: map
22/12/03 10:12:32 INFO mapred.Task: Task 'attempt_local2044251549_0001_m_000000_0' done.
22/12/03 10:12:32 INFO mapred.LocalJobRunner: Finishing task: attempt_local2044251549_0001_m_000000_0
22/12/03 10:12:32 INFO mapred.LocalJobRunner: map task executor complete.
22/12/03 10:12:32 INFO mapreduce.Job: map 100% reduce 0%
22/12/03 10:12:33 INFO mapreduce.Job: Job job_local2044251549_0001 completed successfully
22/12/03 10:12:33 INFO mapreduce.Job: Counters: 20
  File System Counters
    FILE: Number of bytes read=7020
    FILE: Number of bytes written=509834
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=0
    HDFS: Number of bytes written=105
    HDFS: Number of read operations=4
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=3
  Map-Reduce Framework
    Map input records=5
    Map output records=5
    Input split bytes=87
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=29
    Total committed heap usage (bytes)=116391936
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=105
22/12/03 10:12:33 INFO mapreduce.ImportJobBase: Transferred 105 bytes in 5.7622 seconds (18.2221 bytes/sec)
22/12/03 10:12:33 INFO mapreduce.ImportJobBase: Retrieved 5 records.
```

below are the steps/command to see if the data is import to HDFS.

**Commands:** - hadoop fs -ls

```
(hadoopusr@kali)-[~/hadoop-2.9.1]
└─$ hadoop fs -ls
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 10:16:25 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 3 items
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 06:54 Persons
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 10:12 Product
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 07:28 Student
```

**Command:** - Hadoop fs -ls Product

```
(hadoopusr@kali):~/hadoop-2.9.1]$ hadoop fs -ls Product
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: All illegal access operations will be denied in a future release
22/12/03 10:17:16 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r-- 1 hadoopusr supergroup 0 2022-12-03 10:12 Product/_SUCCESS
-rw-r--r-- 1 hadoopusr supergroup 105 2022-12-03 10:12 Product/part-m-00000
```

**Command:** - Hadoop fs -cat Product/part-m-00000

```
(hadoopusr@kali):~/hadoop-2.9.1]$ hadoop fs -cat Product/part-m-00000
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 10:17:37 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
1001,Table,10000,1000
1002,Chair,2000,200
1003,Laptop,20000,2000
1004,Mobile,7000,700
1005,TV,20000,2000
```

2. To Import data in target directory. sqoop

```
import <----- tool
--connect jdbc:mysql://Localhost /database name
--username root
-P
--table Table name
--target-dir filepath<----- HDFS destination dir
```

**Command:** -

```
bin/sqoop import --connect jdbc:mysql://localhost/tsest2 --username root -password kali -
table Product --target-dir /user/hadoopusr/Product2 -m1
```

```
(hadoopusr@kali):[~/sqoop-1.4.7-bin_hadoop-2.6.0]
└─$ bin/sqoop import --connect jdbc:mysql://localhost/tsest2 --username root -password kali --table Product --target-dir /user/hadoopusr/Product2 -m1
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../../zookeeper does not exist! Zookeeper imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
22/12/03 10:28:32 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
22/12/03 10:28:32 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/12/03 10:28:32 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/12/03 10:28:32 INFO tool.CodeGenTool: Beginning code generation
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via its SPI and manual loading of the driver class is generally unnecessary.
22/12/03 10:28:33 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Product` AS t LIMIT 1
22/12/03 10:28:33 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Product` AS t LIMIT 1
22/12/03 10:28:33 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/hadoopusr/hadoop-2.9.1
Note: /tmp/sqoop-hadoopusr/compile/11f7c10bd6c3869fa59b66bdb45ad8f/Product.java uses or overrides a deprecated API.
Note: Recompiling with -Xlint:deprecation for details.
22/12/03 10:28:39 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoopusr/compile/11f7c10bd6c3869fa59b66bdb45ad8f/Product.jar
22/12/03 10:28:39 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/12/03 10:28:39 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/12/03 10:28:39 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/12/03 10:28:39 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/12/03 10:28:39 INFO mapreduce.ImportJobBase: Beginning import of Product
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 10:28:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/12/03 10:28:40 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/12/03 10:28:42 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
22/12/03 10:28:42 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
```

```
Product2/_temporary/0/task_local1680430726_0001_m_000000
22/12/03 10:28:45 INFO mapred.LocalJobRunner: map
22/12/03 10:28:45 INFO mapred.Task: Task 'attempt_local1680430726_0001_m_000000_0' done.
22/12/03 10:28:45 INFO mapred.LocalJobRunner: Finishing task: attempt_local1680430726_0001_m_000000_0
22/12/03 10:28:45 INFO mapred.LocalJobRunner: map task executor complete.
22/12/03 10:28:45 INFO mapreduce.Job: Job job_local1680430726_0001 running in uber mode : false
22/12/03 10:28:45 INFO mapreduce.Job: map 100% reduce 0%
22/12/03 10:28:45 INFO mapreduce.Job: Job job_local1680430726_0001 completed successfully
22/12/03 10:28:45 INFO mapreduce.Job: Counters: 20
  File System Counters
    FILE: Number of bytes read=7020
    FILE: Number of bytes written=509836
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=0
    HDFS: Number of bytes written=105
    HDFS: Number of read operations=4
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=3
  Map-Reduce Framework
    Map input records=5
    Map output records=5
    Input split bytes=87
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=0
    Total committed heap usage (bytes)=141557760
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=105
22/12/03 10:28:45 INFO mapreduce.ImportJobBase: Transferred 105 bytes in 3.506 seconds (29.949 bytes/sec)
22/12/03 10:28:45 INFO mapreduce.ImportJobBase: Retrieved 5 records.
```

**Command:** - hadoop fs -ls

```
(hadoopusr㉿kali)-[~/hadoop-2.9.1]
└─$ hadoop fs -ls
\picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication
b/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflec
WARNING: All illegal access operations will be denied in a future release
22/12/03 10:29:56 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
Found 4 items
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 06:54 Persons
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 10:12 Product
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 10:28 Product2
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 07:28 Student
```

**Command:** - hadoop fs -ls Product2

```
(hadoopusr㉿kali)-[~/hadoop-2.9.1]
└─$ hadoop fs -ls Product2
\picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/li
b/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 10:30:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--  1 hadoopusr supergroup      0 2022-12-03 10:28 Product2/_SUCCESS
-rw-r--r--  1 hadoopusr supergroup  105 2022-12-03 10:28 Product2/part-m-00000
```

**Command:** - hadoop fs -cat Product2/part-m-00000

```
(hadoopusr㉿kali)-[~/hadoop-2.9.1]
└─$ hadoop fs -cat Product2/part-m-00000
\picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/li
b/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 10:30:36 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
1001,Table,10000,1000
1002,Chair,2000,200
1003,Laptop,20000,2000
1004,Mobile,7000,700
1005,TV,20000,2000
```

### 3. Where

```
sqoop import <----- tool
--connect jdbc:mysql://localhost /Database
--username root
-P
--table Table name
--where "productID < 1003"
--target-dir Filepath <---- HDFS destination dir
```

**Command:** -

```
bin/sqoop import --connect jdbc:mysql://localhost/tsest2 --username root -password kali -
table Product --where "P_id < 1002" --target-dir /user/hadoopusr/Product3 -m1.
```

```
(hadoopusr@kali)-[~/sqoop-1.4.7.bin_hadoop-2.6.0]
└─$ bin/sqoop import --connect jdbc:mysql://localhost/ttest2 --username root -password kali --table Product --where "P_id < 1002" --target-dir /user/hadoop
usr/Product3 -m1
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
22/12/03 10:52:18 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
22/12/03 10:52:18 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/12/03 10:52:18 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via its SPI and manual loading of the driver class is generally unnecessary.
22/12/03 10:52:19 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Product` AS t LIMIT 1
22/12/03 10:52:19 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Product` AS t LIMIT 1
22/12/03 10:52:19 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/hadoopusr/hadoop-2.9.1
Note: /tmp/sqoop-hadoopusr/compile/8dbaa0611c48504828b40b4d5291b6d0/Product.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/12/03 10:52:25 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoopusr/compile/8dbaa0611c48504828b40b4d5291b6d0/Product.jar
22/12/03 10:52:25 WARN manager.MySQLManager: It looks like you are importing from mysql.
22/12/03 10:52:25 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/12/03 10:52:25 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/12/03 10:52:25 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/12/03 10:52:25 INFO mapreduce.ImportJobBase: Beginning import of Product
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
```

```
22/12/03 10:52:31 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1056239689_0001_m_000000_0' to hdfs://localhost:54310/user/hadoopusr/Products/_temporary/0/task_local1056239689_0001_m_000000
22/12/03 10:52:31 INFO mapred.LocalJobRunner: map
22/12/03 10:52:31 INFO mapred.Task: Task 'attempt_local1056239689_0001_m_000000_0' done.
22/12/03 10:52:31 INFO mapred.LocalJobRunner: Finishing task: attempt_local1056239689_0001_m_000000_0
22/12/03 10:52:31 INFO mapred.LocalJobRunner: map task executor complete.
22/12/03 10:52:31 INFO mapreduce.Job: Job job_local1056239689_0001 running in uber mode : false
22/12/03 10:52:31 INFO mapreduce.Job: map 100% reduce 0%
22/12/03 10:52:31 INFO mapreduce.Job: Job job_local1056239689_0001 completed successfully
22/12/03 10:52:31 INFO mapreduce.Job: Counters: 20
File System Counters
FILE: Number of bytes read=7020
FILE: Number of bytes written=510261
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=0
HDFS: Number of bytes written=22
HDFS: Number of read operations=4
HDFS: Number of large read operations=0
HDFS: Number of write operations=3
Map-Reduce Framework
Map input records=1
Map output records=1
Input split bytes=87
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=0
Total committed heap usage (bytes)=135266304
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=22
22/12/03 10:52:31 INFO mapreduce.ImportJobBase: Transferred 22 bytes in 3.6259 seconds (6.0675 bytes/sec)
22/12/03 10:52:31 INFO mapreduce.ImportJobBase: Retrieved 1 records.
```

**Command:** - hadoop fs -ls

```
(hadoopusr@kali)-[~/hadoop-2.9.1]
└─$ hadoop fs -ls
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 10:53:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using Java JNA interface instead, which is less efficient
Found 5 items
drwxr-xr-x - hadoopusr supergroup 0 2022-12-03 06:54 Persons
drwxr-xr-x - hadoopusr supergroup 0 2022-12-03 10:12 Product
drwxr-xr-x - hadoopusr supergroup 0 2022-12-03 10:28 Product2
drwxr-xr-x - hadoopusr supergroup 0 2022-12-03 10:52 Product3
drwxr-xr-x - hadoopusr supergroup 0 2022-12-03 07:28 Student
```

**Command:** - hadoop fs -ls Product3

```
(hadoopusr@kali)-[~/hadoop-2.9.1]
$ hadoop fs -ls Product3
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (in unnamed module @0x10000000) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 10:53:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using Java API
Found 2 items
-rw-r--r-- 1 hadoopusr supergroup 0 2022-12-03 10:52 Product3/_SUCCESS
-rw-r--r-- 1 hadoopusr supergroup 22 2022-12-03 10:52 Product3/part-m-00000
```

**Command:** - hadoop fs -cat Product3/part-m-00000

```
(hadoopusr@kali)-[~/hadoop-2.9.1]
$ hadoop fs -cat Product3/part-m-00000
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (in unnamed module @0x10000000) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 10:53:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using Java API
1001,Table,10000,1000
```

#### 4. Query

```
sqoop import <----- tool
--connect jdbc:mysql://Localhost/Database name
--username root
-P
--query 'select * from products WHERE $CONDITIONS' <--- import using query
--target-dir FilePath <----- HDFS destination dir
```

**Command:** -

```
bin/sqoop import --connect jdbc:mysql://localhost/tsest2 --username root -password kali -
query 'select * from Product where P_id> 1002 AND $CONDITIONS' --target-dir
/usr/hadoopusr/Product4 -m1
```

```

(hadoopusr@kali:[~/scoop-1.4.7.bin_hadoop-2.6.0]
$ bin/scoop import --connect jdbc:mysql://localhost/ttest2 -username root -password kali --query 'select * from Product where $CONDITIONS' --split-by p_id --target-dir /user/hadoopusr/Products4 -l
Warning: /home/hadoopusr/scoop-1.4.7.bin_hadoop-2.6.0/bin/ ./hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/hadoopusr/scoop-1.4.7.bin_hadoop-2.6.0/bin/./..../catalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/hadoopusr/scoop-1.4.7.bin_hadoop-2.6.0/bin/./..../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/hadoopusr/scoop-1.4.7.bin_hadoop-2.6.0/bin/./..../zookeeper does not exist! Zookeeper imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Picked up JAVA OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
22/12/03 12:08:54 INFO scoop.Scoop: Running SQuoop version: 1.4.7
22/12/03 12:08:54 WARN tool.BaseSQuoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/12/03 12:08:55 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/12/03 12:08:55 INFO org.CompilationManager: Beginning code generation
Loading class com.mysql.jdbc.Driver. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.
22/12/03 12:08:56 INFO manager.SqlManager: Executing SQL statement: select * from Product where (1 = 0)
22/12/03 12:08:56 INFO manager.SqlManager: Executing SQL statement: select * from Product where (1 = 0)
22/12/03 12:08:56 INFO manager.SqlManager: Executing SQL statement: select * from Product where (1 = 0)
22/12/03 12:08:56 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/hadoopusr/hadoop-2.9.1
Note: Recompile with -Xlint:deprecation for details.
22/12/03 12:09:03 INFO orm.CompilationManager: Writing jar file: /tmp/scoop-hadoopusr/compile/b37ffb919eaff58193d0c4ea3514af5e/QueryResult.jar
22/12/03 12:09:03 INFO mapreduce.ImportJobBase: Beginning query import.
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 12:09:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable
22/12/03 12:09:03 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/12/03 12:09:05 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address

562889272_0001_m_000000
22/12/03 12:09:09 INFO mapred.LocalJobRunner: map
22/12/03 12:09:09 INFO mapred.Task: Task 'attempt_local562889272_0001_m_000000_0' done.
22/12/03 12:09:09 INFO mapred.LocalJobRunner: Finishing task: attempt_local562889272_0001_m_000000_0
22/12/03 12:09:09 INFO mapred.LocalJobRunner: map task executor complete.
22/12/03 12:09:10 INFO mapreduce.Job: map 100% reduce 0%
22/12/03 12:09:10 INFO mapreduce.Job: Job job_local562889272_0001 completed successfully
22/12/03 12:09:10 INFO mapreduce.Job: Counters: 20
    File System Counters
        FILE: Number of bytes read=7108
        FILE: Number of bytes written=507699
        FILE: Number of read operations=0
        FILE: Number of large read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=0
        HDFS: Number of bytes written=105
        HDFS: Number of read operations=4
        HDFS: Number of large read operations=0
        HDFS: Number of write operations=3
    Map-Reduce Framework
        Map input records=5
        Map output records=5
        Input split bytes=87
        Spilled Records=0
        Failed Shuffles=0
        Merged Map outputs=0
        GC time elapsed (ms)=35
        Total committed heap usage (bytes)=118489088
    File Input Format Counters
        Bytes Read=0
    File Output Format Counters
        Bytes Written=105
22/12/03 12:09:10 INFO mapreduce.ImportJobBase: Transferred 105 bytes in 4.6414 seconds (22.6226 bytes/sec)
22/12/03 12:09:10 INFO mapreduce.ImportJobBase: Retrieved 5 records.

```

**Command:** - hadoop fs -ls

```

(hadoopusr@kali:[~/hadoop-2.9.1]
$ hadoop fs -ls
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 12:09:42 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable
Found 6 items
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 06:54 Persons
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 10:12 Product
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 10:28 Product2
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 10:52 Product3
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 12:09 Product4
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 07:28 Student

```

**Command:** - hadoop fs -ls Product4

```
(hadoopusr@kali)-[~/hadoop-2.9.1]
$ hadoop fs -ls Product4
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil$KerberosConfig.getINSTANCE()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 12:10:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin Java classes where applicable
Found 2 items
-rw-r--r-- 1 hadoopusr supergroup          0 2022-12-03 12:09 Product4/_SUCCESS
-rw-r--r-- 1 hadoopusr supergroup      105 2022-12-03 12:09 Product4/part-m-00000
```

**Command:** - hadoop fs -cat Product4/part-m-00000

```
(hadoopusr@kali)-[~/hadoop-2.9.1]
$ hadoop fs -cat Product4/part-m-00000
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil$KerberosConfig.getINSTANCE()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 12:10:34 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin Java classes where applicable
1001,Table,10000,1000
1002,Chair,2000,200
1003,Laptop,20000,2000
1004,Mobile,7000,700
1005,TV,20000,2000
```

## 5. Query with WHERE

```
sqoop import <----- tool
--connect jdbc:mysql://Localhost /Database name
--username root
-P
--query 'select * from products WHERE productID > 1003 AND $CONDITIONS' <--- import subset of query
--split-by productID
--target-dir Filepath<----- HDFS destination dir
```

**Command: -**

```
bin/sqoop import --connect jdbc:mysql://localhost/tsest2 --username root -password kali -  
query 'select * from Product where P_id> 1002 AND $CONDITIONS' --target-dir  
/user/hadoopusr/Product5 -m1
```

```

[hadoopusr@kali]:~/sqoop-1.4.7-bin_hadoop-2.6.0]
$ bin/sqoop import --connect jdbc:mysql://localhost/ttest2 --username root -password kali --query 'select * from Product where P_id> 1002 AND $CONDITIONS' --target-dir /user/hadoopusr/
Product5 -m1
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../zookeeper does not exist! Zookeeper imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Picked up JAVA_TOOL_OPTIONS: -Djava.security.krb5.conf=/etc/krb5.conf -Djava.security.auth.login.config=/etc/java-security/auth-login.conf
22/12/03 12:14:47 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
22/12/03 12:14:47 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/12/03 12:14:47 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/12/03 12:14:47 INFO tool.CodeGenTool: Beginning code generation
Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.
22/12/03 12:14:47 INFO manager.SqlManager: Executing SQL statement: select * from Product where P_id> 1002 AND (1 = 0)
22/12/03 12:14:48 INFO manager.SqlManager: Executing SQL statement: select * from Product where P_id> 1002 AND (1 = 0)
22/12/03 12:14:48 INFO manager.SqlManager: Executing SQL statement: select * from Product where P_id> 1002 AND (1 = 0)
22/12/03 12:14:48 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/hadoopusr/hadoop-2.9.1
Note: /tmp/sqoop-hadoopuser/compile/03c71cbafee61fc2106f28138c6f266/QueryResult.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/12/03 12:14:50 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoopuser/compile/03c71cbafee61fc2106f28138c6f266/QueryResult.jar
22/12/03 12:14:50 INFO mapreduce.Job:  map 0% reduce 0%
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getINSTANCE()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use -illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 12:14:55 WARN Configuration.deprecation: mapred.lib.to hadoop-mapreduce library for your platform... using builtin-Java classes where applicable
22/12/03 12:14:57 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/12/03 12:14:57 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
22/12/03 12:14:57 INFO Configuration.deprecation: session.job.id is deprecated. Instead, use dfs.metrics.session-id
22/12/03 12:14:57 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
22/12/03 12:14:57 INFO db.DBInputFormat: Using read committed transaction isolation

```

```

22/12/03 12:15:00 INFO output.FileOutputCommitter: Saved output of task 'attempt_local1253460493_0001_m_000000_0'
l1253460493_0001_m_000000
22/12/03 12:15:00 INFO mapred.LocalJobRunner: map
22/12/03 12:15:00 INFO mapred.Task: Task 'attempt_local1253460493_0001_m_000000_0' done.
22/12/03 12:15:00 INFO mapred.LocalJobRunner: Finishing task: attempt_local1253460493_0001_m_000000_0
22/12/03 12:15:00 INFO mapred.LocalJobRunner: map task executor complete.
22/12/03 12:15:00 INFO mapreduce.Job: Job job_local1253460493_0001 running in uber mode : false
22/12/03 12:15:00 INFO mapreduce.Job: map 100% reduce 0%
22/12/03 12:15:01 INFO mapreduce.Job: Job job_local1253460493_0001 completed successfully
22/12/03 12:15:01 INFO mapreduce.Job: Counters: 20
File System Counters
  FILE: Number of bytes read=7108
  FILE: Number of bytes written=509602
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=0
  HDFS: Number of bytes written=63
  HDFS: Number of read operations=4
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=3
Map-Reduce Framework
  Map input records=3
  Map output records=3
  Input split bytes=87
  Spilled Records=0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=15
  Total committed heap usage (bytes)=135266304
File Input Format Counters
  Bytes Read=0
File Output Format Counters
  Bytes Written=63
22/12/03 12:15:01 INFO mapreduce.ImportJobBase: Transferred 63 bytes in 4.7521 seconds (13.2573 bytes/sec)
22/12/03 12:15:01 INFO mapreduce.ImportJobBase: Retrieved 3 records.

```

### Command: -

hadoop fs -ls Product5

```
(hadoopusr@kali)-[~/hadoop-2.9.1]
$ hadoop fs -ls
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil$KerberosConfig.getConf() from class org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 12:15:22 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin Java classes where applicable
Found 7 items
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 06:54 Persons
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 10:12 Product
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 10:28 Product2
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 10:52 Product3
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 12:09 Product4
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 12:15 Product5
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 07:28 Student
```

**Command:** - hadoop fs -ls Product5

```
(hadoopusr@kali)-[~/hadoop-2.9.1]
$ hadoop fs -ls Product5
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil$KerberosConfig.getConf() from class org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 12:15:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin Java classes where applicable
Found 2 items
-rw-r--r--  1 hadoopusr supergroup          0 2022-12-03 12:15 Product5/_SUCCESS
-rw-r--r--  1 hadoopusr supergroup         63 2022-12-03 12:15 Product5/part-m-00000
```

**Command:** - hadoop fs -cat Product5/part-m-00000

```
(hadoopusr@kali)-[~/hadoop-2.9.1]
$ hadoop fs -cat Product5/part-m-00000
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil$KerberosConfig.getConf() from class org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 12:16:13 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin Java classes where applicable
1003,Laptop,20000,2000
1004,Mobile,7000,700
1005,TV,20000,2000
```

## 6. Incremental Import

**Command:** - bin/sqoop  
import \

```
--connect jdbc:mysql://localhost/ttest2 \
--username root \
-P \
--table Product \
--m 1 \
--incremental append \
--check-column P_id \
--last-value 1005
```

```
(hadoopusr@kali:[~/sqoop-1.4.7.bin_hadoop-2.6.0]
└─$ bin/sqoop import \
  --connect jdbc:mysql://localhost/ttest2 \
  --username root \
  -P \
  --table Product \
  --m 1 \
  --incremental append \
  --check-column P_id \
  --last-value 1005
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../zookeeper does not exist! Zookeeper imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
22/12/04 09:54:48 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
22/12/04 09:54:52 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/12/04 09:54:52 INFO tool.CodeGenTool: Beginning code generation
Loading class 'com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI and manual loading of the driver class is generally unnecessary.
22/12/04 09:54:53 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Product` AS t LIMIT 1
22/12/04 09:54:53 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Product` AS t LIMIT 1
22/12/04 09:54:53 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/hadoopusr/hadoop-2.9.1
Note: /tmp/sqoop-hadoopusr/compile/ee58ce6b3baa15ed33ac5f67642941a/Product.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/12/04 09:55:00 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoopusr/compile/ee58ce6b3baa15ed33ac5f67642941a/Product.jar
22/12/04 09:55:00 INFO tool.ImportTool: Maximal id query for free form incremental import: SELECT MAX(`P_id`) FROM `Product`
22/12/04 09:55:00 INFO tool.ImportTool: Incremental import based on column `P_id`
22/12/04 09:55:00 INFO tool.ImportTool: Lower bound value: 1005
22/12/04 09:55:00 INFO tool.ImportTool: Upper bound value: 1006
```

```
(hadoopusr@kali:[~/hadoop-2.9.1]
└─$ hadoop fs -cat Product/part-m-00001
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use -illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/04 09:57:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
1006,AC,20000,2000
```

## 7. Import all table.

**Command:** -

```
bin/sqoop import-all-tables --connect jdbc:mysql://localhost/test2 --username root
password kali -m1
```

```

[hadoopusr@kali] -~/sqoop-1.4.7-bin_hadoop-2.6.0]
$ bin/sqoop import-all-tables --connect jdbc:mysql://localhost/test2 --username root -password kali -m1
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../../zookeeper does not exist! Zookeeper imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=true -Dswing.aatext=true
22/12/03 12:55:37 INFO sqoop: Sqoop: Running Sqoop version: 1.4.7
22/12/03 12:55:37 WARN tool:BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/12/03 12:55:38 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
Loading class com.mysql.jdbc.Driver. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered via the SPI e driver is generally unnecessary.
22/12/03 12:55:39 INFO tool.CodeGenTool: Beginning code generation
22/12/03 12:55:39 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `demo2` AS t LIMIT 1
22/12/03 12:55:39 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `demo2` AS t LIMIT 1
22/12/03 12:55:39 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/hadoopusr/hadoop-2.9.1
Note: /tmp/sqoop-hadoopusr/compile/d0063b21171a272eab1d1645ad6e2e9/demo2.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/12/03 12:55:46 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoopusr/compile/d0063b21171a272eab1d1645ad6e2e9/demo2.jar
22/12/03 12:55:46 WARN manager.MySQLManager: MySQL looks like you are importing from mysql.
22/12/03 12:55:46 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
22/12/03 12:55:46 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
22/12/03 12:55:46 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
22/12/03 12:55:46 INFO oracleImportJob:Beginning import of demo2
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop_sun_security_krb5_Config.getInstance())
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 12:55:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/12/03 12:55:47 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/12/03 12:55:49 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address

```

hadoop fs -ls

```

[hadoopusr@kali] -~/hadoop-2.9.1]
$ hadoop fs -ls
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=true -Dswing.aatext
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop_sun_security_krb5_Config.getInstance())
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 14:42:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 13 items
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 06:54 Persons
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 10:12 Product
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 10:28 Product2
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 10:52 Product3
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 12:09 Product4
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 12:15 Product5
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 12:38 Product6
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 12:45 Product7
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 07:28 Student
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 12:55 demo1
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 12:55 demo2
drwxr-xr-x  - hadoopusr supergroup          0 2022-12-03 12:56 demo3
-rw-r--r--  1 hadoopusr supergroup          14 2022-12-03 13:27 export

```

## 8. Eval –Query

**Command:** -

```
bin/sqoop eval --connect jdbc:mysql://localhost/tsest2 --username root -P --query "select * from Product"
```

```
(hadoopusr㉿kali)-[~/sqoop-1.4.7.bin_hadoop-2.6.0]
└─$ bin/sqoop eval --connect jdbc:mysql://localhost/tsest2 --username root -P
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../hbase does n
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../hcatalog doe
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../accumulo doe
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../zookeeper do
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
22/12/04 11:12:44 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
22/12/04 11:12:47 INFO manager.MySQLManager: Preparing to use a MySQL streaming
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver cla
he SPI and manual loading of the driver class is generally unnecessary.

+-----+-----+-----+-----+
| P_id | P_Name | P_Prices | P_Discount |
+-----+-----+-----+-----+
| 1001 | Table  | 10000   | 1000     |
| 1002 | Chair   | 2000    | 200      |
| 1003 | Laptop  | 20000   | 2000     |
| 1004 | Mobile  | 7000    | 700      |
| 1005 | TV      | 20000   | 2000     |
| 1006 | AC      | 20000   | 2000     |
+-----+-----+-----+-----+
```

## 9. EVAL - -Insert

```
bin/sqoop eval --connect jdbc:mysql://localhost/tsest2 --username root -P -e "INSERT
INTO Product VALUES(1009,'Watches',20000,2000);"
```

```
(hadoopusr㉿kali)-[~/sqoop-1.4.7.bin_hadoop-2.6.0]
└─$ bin/sqoop eval --connect jdbc:mysql://localhost/tsest2 --username root -P -e "INSERT INTO Product VALUES(1009,'Watches',20000,2000);"
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../zookeeper does not exist! Zookeeper imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
22/12/04 11:14:38 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
22/12/04 11:14:41 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is automatically regi
he SPI and manual loading of the driver class is generally unnecessary.
22/12/04 11:14:42 INFO tool.EvalSqlTool: 1 row(s) updated.
```

## 10. List Database

```
bin/sqoop list-databases --connect jdbc:mysql://localhost/tsest2 --username root -P
```

```
(hadoopusr@kali)-[~/sqoop-1.4.7.bin_hadoop-2.6.0]
└─$ bin/sqoop list-databases --connect jdbc:mysql://localhost/tsest2 --username root -P
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../zookeeper does not exist! Zookeeper imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
22/12/04 11:17:28 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
22/12/04 11:17:31 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is SPI and manual loading of the driver class is generally unnecessary.
information_schema
sys
tsest2
test
mysql
test2
performance_schema
```

## 11. List table.

```
bin/sqoop list-
tables \
--connect jdbc:mysql://localhost/tsest2 \
--username root \
-P
```

```
(hadoopusr@kali)-[~/sqoop-1.4.7.bin_hadoop-2.6.0]
└─$ bin/sqoop list-tables \
--connect jdbc:mysql://localhost/tsest2 \
--username root \
--password kali
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/hadoopusr/sqoop-1.4.7.bin_hadoop-2.6.0/bin/../../zookeeper does not exist! Zookeeper imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
22/12/04 11:20:11 INFO sqoop.Sqoop: Running Sqoop version: 1.4.7
Enter password:
22/12/04 11:20:14 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is `com.mysql.cj.jdbc.Driver'. The driver is SPI and manual loading of the driver class is generally unnecessary.
Product
```

## 12. Export data from HDFS to MySql.

**Command:** -

```
bin/sqoop export --connect jdbc:mysql://localhost/test2 --username root -password kali -
table export --export-dir /user/hadoopusr/export
```

```

[hadoopusr@kali:~/sqoop-1.4.7-bin_hadoop-2.6.0]
$ bin/sqoop export --connect jdbc:mysql://localhost/test2 --username root -password kali --table export --export-dir /user/hadoopusr/export
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /home/hadoopusr/sqoop-1.4.7-bin_hadoop-2.6.0/bin/../../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
22/12/03 13:33:02 INFO sqoop: Running Sqoop version: 1.4.7
22/12/03 13:33:02 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/12/03 13:33:02 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
22/12/03 13:33:02 INFO tool.CodeGenTool: Beginning code generation
Loading class `com.mysql.jdbc.Driver'. This is deprecated. The new driver class is 'com.mysql.cj.jdbc.Driver'. The driver is automatically registered
e driver class is generally unnecessary.
22/12/03 13:33:04 INFO manager.SQLManager: Executing SQL statement: SELECT t.* FROM `export` AS t LIMIT 1
22/12/03 13:33:04 INFO manager.SQLManager: Executing SQL statement: SELECT t.* FROM `export` AS t LIMIT 1
22/12/03 13:33:04 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /home/hadoopusr/hadoop-2.9.1
Note: /tmp/sqoop-hadoopusr/compile/67b4a4944c0cd5eb7fcccfa1a3dd8f/export.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
22/12/03 13:33:13 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-hadoopusr/compile/67b4a4944c0cd5eb7fcccfa1a3dd8f/export.jar
22/12/03 13:33:15 INFO mapreduce.ExportJobBase: Beginning export of export
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/comm
hod.sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/03 13:33:15 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
22/12/03 13:33:14 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
22/12/03 13:33:19 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.speculative
22/12/03 13:33:19 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculative
22/12/03 13:33:19 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
22/12/03 13:33:20 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
22/12/03 13:33:20 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=

```

```

22/12/03 13:34:07 INFO mapred.Task: Task:attempt_local168324352_0001_m_000003_0 is done. And is in the process of committing
22/12/03 13:34:07 INFO mapred.LocalJobRunner: map
22/12/03 13:34:07 INFO mapred.Task: Task:attempt_local168324352_0001_m_000003_0' done.
22/12/03 13:34:07 INFO mapred.LocalJobRunner: Finishing task: attempt_local168324352_0001_m_000003_0
22/12/03 13:34:07 INFO mapred.LocalJobRunner: map task executor complete.
22/12/03 13:34:08 INFO mapreduce.Job: map 100% reduce 0%
22/12/03 13:34:08 INFO mapreduce.Job: Job job_local168324352_0001 completed successfully
22/12/03 13:34:08 INFO mapreduce.Job: Counters: 20
  File System Counters
    FILE: Number of bytes read=23744
    FILE: Number of bytes written=2019172
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=158
    HDFS: Number of bytes written=0
    HDFS: Number of read operations=66
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=0
  Map-Reduce Framework
    Map input records=7
    Map output records=7
    Input split bytes=561
    Spilled Records=0
    Failed Shuffles=0
    Merged Map outputs=0
    GC time elapsed (ms)=0
    Total committed heap usage (bytes)=654311424
  File Input Format Counters
    Bytes Read=0
  File Output Format Counters
    Bytes Written=0
22/12/03 13:34:08 INFO mapreduce.ExportJobBase: Transferred 158 bytes in 48.1346 seconds (3.2825 bytes/sec)
22/12/03 13:34:08 INFO mapreduce.ExportJobBase: Exported 7 records.

```

```

MariaDB [test2]> select * from export;
+-----+
| e_id |
+-----+
|   6   |
|   7   |
|   1   |
|   2   |
|   3   |
|   4   |
|   5   |
+-----+
7 rows in set (0.006 sec)

```

## Practical 6

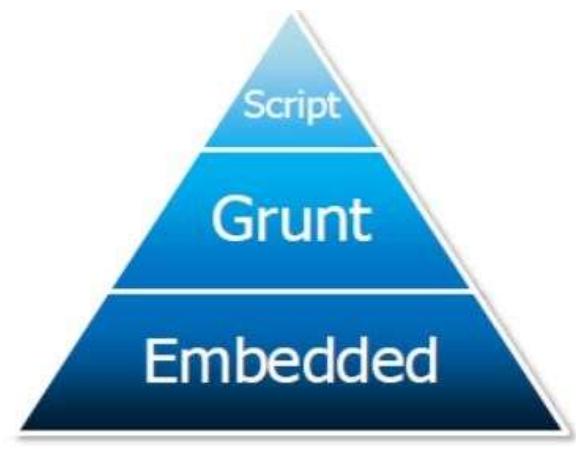
**Aim:** Write a Pig Script for solving counting problems

### What is Apache Pig?

Pig is a high-level programming language useful for analyzing large data sets. Pig was a result of development effort at Yahoo. Apache Pig enables people to focus more on analyzing bulk data sets and to spend less time writing Map-Reduce programs. the Apache Pig programming language is designed to work upon any kind of data.

Pig is important as companies like Yahoo, Google and Microsoft are collecting huge amounts of data sets and is also used in some form of ad-hoc processing and analysis of all the information.

### Basic Program Structure of Pig:



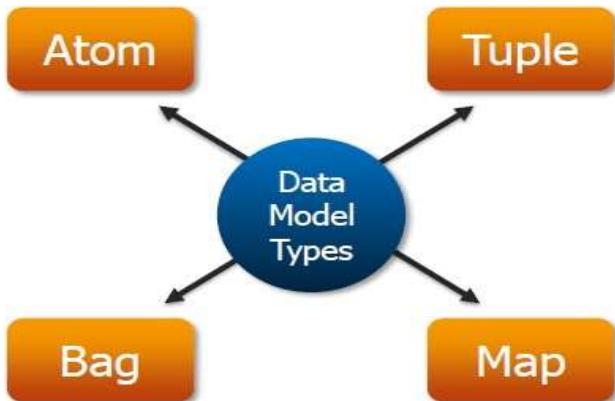
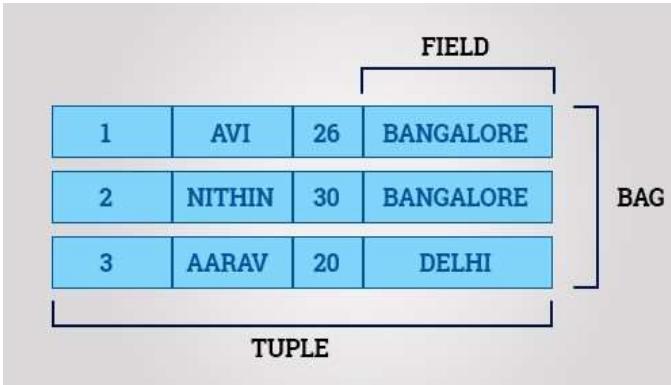
---

Script – Pig Can run a file script that contains Pig Commands. Eg: `pig script .pig` runs the command in the local file `script.pig`.

Grunt – It is an interactive shell for running Pig commands. It is also possible to run pig scripts from within Grunts using `run` and `exec`.

Embedded – Can run Pig programs from Java, much like you can use JDBC to run SQL programs from Java.

### Basic Types of Data Models in Pig:



Pig comprises of 4 basic types of data models. They are as follows:

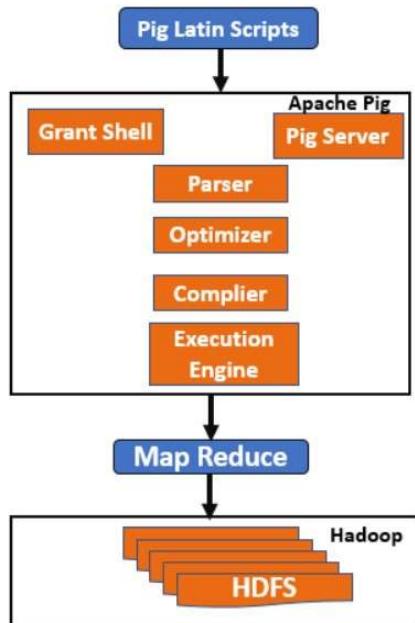
Atom – It is a simple atomic data value. It is stored as a string but can be used as either a string or a number.

Tuple – An ordered set of fields.

Bag – A collection of tuples.

Map – set of key value pairs.

### **Pig Architecture:**



1. **Parser:** When a Pig Latin script is sent to Hadoop Pig, it is first handled by the parser. The parser is responsible for checking the syntax of the script. Parser gives an output in the form of a Directed Acyclic Graph (DAG) that contains Pig Latin statements, together with other logical operators represented as nodes.
2. **Optimizer:** After the output from the parser is retrieved, a logical plan for DAG is passed to a logical optimizer. The optimizer is responsible for carrying out the logical optimizations.
3. **Compiler:** The role of the compiler comes in when the output from the optimizer is received. The compiler compiles the logical plan sent by the optimizer into a series of MapReduce tasks or jobs.
4. **Execution Engine:** After the logical plan is converted to MapReduce jobs, these jobs are sent to Hadoop in a properly sorted order, and these jobs are executed on Hadoop for yielding the desired result.

#### Hadoop Pig Features:

1. Rich set of operators

One of the major advantages is, in order to perform several operations, there is a huge set of operators offered by Apache Pig, such as join, sort, filer, etc.

2. Ease of programming

Basically, for SQL Programmer, Pig Latin is a boon. It is as similar to SQL. Hence, if you are good at SQL it is easy to write a Pig script.

### 3. Optimization opportunities

Also, it's a benefit working here because in Apache Pig the tasks optimize their execution automatically. Hence, as a result, programmers only need to focus on the semantics of the language.

#### **Components of Pig**

There are two major components of the Pig:

- Pig Latin script language
- A runtime engine

Pig Latin script language

The Pig Latin script is a procedural data flow language. It contains syntax and commands that can be applied to implement business logic. Examples of Pig Latin are LOAD and STORE.

A runtime engine

The runtime engine is a compiler that produces sequences of MapReduce programs. It uses HDFS to store and retrieve data. It is also used to interact with the Hadoop system (HDFS and MapReduce).

The runtime engine parses, validates, and compiles the script operations into a sequence of MapReduce jobs.

#### **Apache Pig Installation:**

Pre-Requisite to Install Pig

You must have Hadoop and Java JDK installed on your system.

- **java Installation** - Check whether the Java is installed or not using the following command.

`$java -version` ◦ **Hadoop Installation** - Check whether the Hadoop is installed or not using the following command.

`$Shadoop version`

1. To download pig, we need to go on its website- <https://pig.apache.org/releases.html>



apache pig download



All Videos Images News Shopping More Tools

About 11,60,000 results (0.36 seconds)

<https://pig.apache.org> › Pig

## Apache Pig Releases

Download. Releases may be downloaded from Apache mirrors. Download a release now! Get

Pig .rpm or .deb. Starting with Pig 0.12 ...

Download · News · July, 2014: release 0.13.0...

2. Then click on downloads-



Project Wiki

Project

- Welcome
- Releases
- About
- Mailing Lists
- Who We Are
- Bylaws
- Pig Tools
- Privacy Policy
- Sponsorship
- Thanks

Documentation Developers

APACHECON OCTOBER 3 - 6, 2022 NEW ORLEANS LOUISIANA

## Apache Pig Releases

Download News

- 19 June, 2017: release 0.17.0 available
- 8 June, 2016: release 0.16.0 available
- 6 June, 2015: release 0.15.0 available
- 20 November, 2014: release 0.14.0 available
- 4 July, 2014: release 0.13.0 available
- 14 April, 2014: release 0.12.1 available
- 14 October, 2013: release 0.12.0 available
- 1 April, 2013: release 0.11.1 available
- 21 February, 2013: release 0.11.0 available
- 6 January, 2013: release 0.10.1 available
- 25 April, 2012: release 0.10.0 available
- 22 January, 2012: release 0.9.2 available
- 5 October, 2011: release 0.9.1 available
- 29 July, 2011: release 0.9.0 available
- 24 April, 2011: release 0.8.1 available
- 17 December, 2010: release 0.8.0 available
- 13 May, 2010: release 0.7.0 available
- 1 March, 2010: release 0.6.0 available
- 29 October, 2009: release 0.5.0 available
- 29 September, 2009: release 0.4.0 available
- 25 June, 2009: release 0.3.0 available
- 8 April, 2009: release 0.2.0 available
- 5 December, 2008: release 0.1.1 available
- 11 September, 2008: release 0.1.0 available

3. Now we have to download the required version of pig-



COMMUNITY-LED DEVELOPMENT "THE APACHE WAY"

Projects ▾ People ▾ Community ▾ License ▾ Sponsors ▾



We suggest the following site for your download:

<https://dlcdn.apache.org/pig>

Alternate download locations are suggested below.

It is essential that you verify the integrity of the downloaded file using the PGP signature (.asc file) or a hash (.md5 or .sha\* file).

## HTTP

<https://dlcdn.apache.org/pig>

## BACKUP SITE

<https://dlcdn.apache.org/pig>

o Download

the Apache Pig tar file.

- o Unzip the downloaded tar file.

```
$ tar -xvf pig-0.16.0.tar.gz  o Open the  
bashrc file.
```

```
$ sudo nano ~/.bashrc
```

Now, provide the following PIG\_HOME path.

```
export PIG_HOME=/home/hduser/pig-0.16.0  export  
PATH=$PATH:$PIG_HOME/bin
```

```
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64  
export HADOOP_INSTALL=/home/hadoopusr/hadoop-2.9.1  
export PATH=$PATH:$HADOOP_INSTALL/bin  
export PATH=$PATH:$HADOOP_INSTALL/sbin  
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL  
export HADOOP_COMMON_HOME=$HADOOP_INSTALL  
export HADOOP_HDFS_HOME=$HADOOP_INSTALL  
export YARN_HOME=$HADOOP_INSTALL  
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native  
export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib"  
#HADOOP VARIABLES END  
export PIG_HOME=/home/hadoopusr/pig-0.16.0  
export PATH=$PATH:$PIG_HOME/bin
```

- o Update the environment variable

```
$ source ~/.bashrc  o Let's test the installation on the command prompt type $ pig -  
h
```

Rajeev Katara

o

```
(hadoopusr@kali)-[~/hadoop-2.9.1]
$ pig -h
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true

Apache Pig version 0.16.0 (r1746530)
compiled Jun 01 2016, 23:10:49

USAGE: Pig [options] [-] : Run interactively in grunt shell.
      Pig [options] -e[xcute] cmd [cmd ...] : Run cmd(s).
      Pig [options] [-f[file]] file : Run cmds found in file.
options include:
  -4, -log4jconf - Log4j configuration file, overrides log conf
  -b, -brief - Brief logging (no timestamps)
  -c, -check - Syntax check
  -d, -debug - Debug level, INFO is default
  -e, -execute - Commands to execute (within quotes)
  -f, -file - Path to the script to execute
  -g, -embedded - ScriptEngine classname or keyword for the ScriptEngine
  -h, -help - Display this message. You can specify topic to get help for that topic.
    properties is the only topic currently supported: -h properties.
  -i, -version - Display version information
  -l, -logfile - Path to client side log file; default is current working directory.
  -m, -param_file - Path to the parameter file
  -p, -param - Key value pair of the form param=val
  -r, -dryrun - Produces script with substituted parameters. Script is not executed.
  -t, -optimizer_off - Turn optimizations off. The following values are supported:
    ConstantCalculator - Calculate constants at compile time
    SplitFilter - Split filter conditions
    PushUpFilter - Filter as early as possible
    MergeFilter - Merge filter conditions
    PushDownForEachFlatten - Join or explode as late as possible
    LimitOptimizer - Limit as early as possible
```

- o Let's start the pig in MapReduce mode. \$ pig

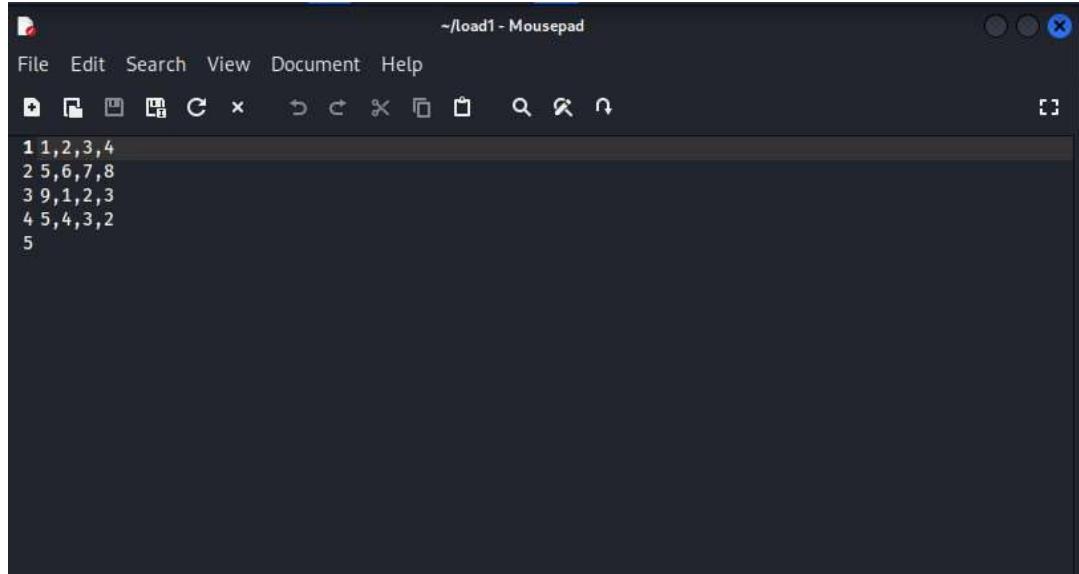
```
(hadoopusr@kali)-[~]
$ pig
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
22/12/06 14:10:51 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
22/12/06 14:10:51 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
22/12/06 14:10:51 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2022-12-06 14:10:51,869 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2022-12-06 14:10:51,869 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoopusr/pig_1670353851863.log
2022-12-06 14:10:51,979 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/hadoopusr/.pigbootup not found
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
2022-12-06 14:10:53,287 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2022-12-06 14:10:53,410 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-12-06 14:10:53,411 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-06 14:10:53,411 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:54310
2022-12-06 14:10:55,295 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PTG-default-46980134-961f-4615-8738-239f4371c4c5
2022-12-06 14:10:55,295 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> ■
```

## Implementation:

### Implementing the pig operators:

#### 1. Load:

Create a text file in your local machine and provide some values to it. Check the values written in the text files.



Upload the text files on HDFS in the specific directory.

```
(hadoopusr@kali)-[~/hadoop-2.9.1]
$ hadoop fs -put /home/hadoopusr/load1 /user/hadoopusr
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.newInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/06 14:21:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

- Open the pig MapReduce run mode.

```
(hadoopusr@kali)-[~]
$ pig
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
22/12/06 14:24:47 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
22/12/06 14:24:47 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
22/12/06 14:24:47 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2022-12-06 14:24:47,856 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2022-12-06 14:24:47,857 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoopusr/pig_1670354687851.log
2022-12-06 14:24:47,943 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/hadoopusr/.pigbootup not found
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
2022-12-06 14:24:48,950 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-classes where applicable
2022-12-06 14:24:49,032 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-12-06 14:24:49,032 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-06 14:24:49,032 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:54310
2022-12-06 14:24:50,684 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-8a18e957-6d38-444c-8066-52b2b57e6679
2022-12-06 14:24:50,685 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> ■
```

- Load the file that contains the data.

grunt> A = LOAD '/pigexample/pload.txt' USING PigStorage(',') AS ( a1:int,a2:int,a3:int,a4:int );

```
grunt> A = LOAD '/user/hadoopusr/load1' USING PigStorage(',') AS ( a1:int,a2:int,a3:int,a4:int );
2022-12-06 14:28:16,841 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> DUMP A
```

- Now, execute and verify the data. grunt> DUMP A;

```

o
2022-12-06 14:28:16,841 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunts> DUMP A
2022-12-06 14:28:23,514 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2022-12-06 14:28:23,538 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-06 14:28:23,539 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-12-06 14:28:23,540 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2022-12-06 14:28:23,544 [main] INFO org.apache.pig.backend.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic
? false
2022-12-06 14:28:23,546 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-12-06 14:28:23,546 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-12-06 14:28:23,562 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-06 14:28:23,573 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2022-12-06 14:28:23,576 [main] INFO org.apache.pig.tools.pigstats.mapreduce.MRScriptState - Pig script settings are added to the job
2022-12-06 14:28:23,577 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.p
2022-12-06 14:28:31,183 [main]
2022-12-06 14:28:31,184 [main]
(1,2,3,4)
(5,6,7,8)
(9,1,2,3)
(5,4,3,2)

```

- Let's check the corresponding schema. grunt> DESCRIBE A;

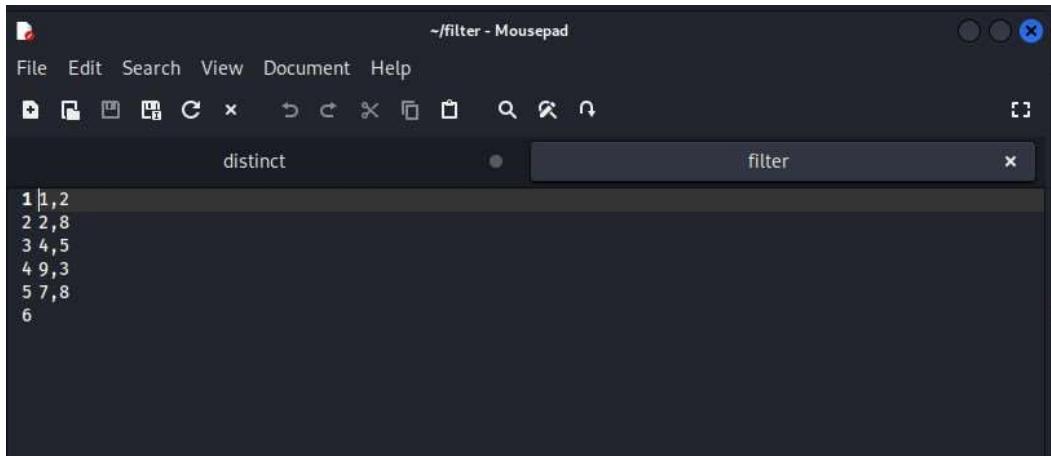
```

(5,4,3,2)
grunt> DESCRIBE A;
A: {a1: int,a2: int,a3: int,a4: int}
grunt> 

```

## 2. Filter:

Create a text file in your local machine and provide some values to it.



Upload the text files on HDFS in the specific directory.

```

[hadoopusr@Kali:~/hadoop-2.9.1]
└─$ hadoop fs -put /home/hadoopusr/filter /user/hadoopusr
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/06 14:41:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

```

Open the pig MapReduce run mode.

\$ pig

Rajeev Katara

CS21006

```

○
└─(hadoopusr㉿kali) [~]
$ pig
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
22/12/06 14:24:47 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
22/12/06 14:24:47 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
22/12/06 14:24:47 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2022-12-06 14:24:47,856 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2022-12-06 14:24:47,857 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoopusr/pig_1670354687851.log
2022-12-06 14:24:47,943 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/hadoopusr/.pigbootup not found
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
2022-12-06 14:24:48,950 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using builtin-classes where applicable
2022-12-06 14:24:49,032 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-12-06 14:24:49,032 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-06 14:24:49,032 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:54310
2022-12-06 14:24:50,684 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-8a18e957-6d38-444c-8066-52b2b57e6679
2022-12-06 14:24:50,685 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> ■

```

- Load the file that contains the data.

grunt> A = LOAD '/pigexample/pfilter.txt' USING PigStorage(',') AS (a1:int,a2:int);

```

grunt> A = LOAD '/user/hadoopusr/filter' USING PigStorage(',') AS (a1:int,a2:int);
2022-12-06 14:47:58,736 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> Dump A;
2022-12-06 14:48:09,282 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2022-12-06 14:48:09,318 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-06 14:48:09,318 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-12-06 14:48:09,319 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2022-12-06 14:48:09,324 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic
? false
2022-12-06 14:48:09,327 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-12-06 14:48:09,327 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-12-06 14:48:09,356 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-06 14:48:09,360 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alr

```

○

Now, execute and verify the data grunt> DUMP A;

```

grunt> A = LOAD '/user/hadoopusr/filter' USING PigStorage(',') AS (a1:int,a2:int);
2022-12-06 14:47:58,736 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> Dump A;
2022-12-06 14:48:09,282 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2022-12-06 14:48:09,318 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-06 14:48:09,318 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-12-06 14:48:09,319 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2022-12-06 14:48:09,324 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic
? false
2022-12-06 14:48:09,327 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 1
2022-12-06 14:48:09,327 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-12-06 14:48:09,356 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-06 14:48:09,360 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alr
2022-12-06 14:48:19,350 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2022-12-06 14:48:19,357 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2022-12-06 14:48:19,365 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2022-12-06 14:48:19,396 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-12-06 14:48:19,402 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-06 14:48:19,402 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-06 14:48:19,424 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-06 14:48:19,425 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,2)
(2,8)
(4,5)
(9,3)
(7,8)

```

Let's execute FILTER operator to eliminate duplicate tuples.

grunt> Result = FILTER A BY a2==8;

```

grunt> Result = FILTER A BY a2==8;
grunt> Dump Result;

```

- Now, execute and verify the data. grunt> DUMP Result;

Rajeev Katara

```

grunt> dump Result;
2022-12-06 14:48:47,807 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: FILTER
2022-12-06 14:48:47,807 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-06 14:48:47,831 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-06 14:48:47,832 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - [RULES_ENABLED=AddForEach, ColumnMapKeyPrune, ConstantTunneling, GroupByForParallelFilter, LimitOptimizer, MapReduceOptimizer, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimize, PushdownForJoin, PushdownForLeftOuterJoin, StreamCastOptimizer]
2022-12-06 14:48:47,839 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic
? false
2022-12-06 14:48:47,844 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 4
2022-12-06 14:48:47,844 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 1
2022-12-06 14:48:47,859 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-06 14:48:47,861 [main] INFO org.apache.hadoop.metrics.jvm.JVMMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - alr
eady initialized
2022-12-06 14:48:47,866 [main] INFO org.apache.pig.tools.pigstats.MWScriptState - Pig script settings are added to the job
2022-12-06 14:48:47,866 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2022-12-06 14:48:47,867 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - This job cannot be converted run in-process
2022-12-06 14:48:47,140 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/hadoopusr/pig-0.13.1.jar to DistributedCache through /tmp/11836216/tmp-113822416/automaton-1.11-8.jar
2022-12-06 14:48:47,592 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/hadoopusr/pig-0.16.0/lib/automaton-1.11-8.jar to DistributedCache through /tmp/11836216/tmp-113822416/automaton-1.11-8.jar
2022-12-06 14:48:47,819 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/hadoopusr/pig-0.16.0/lib/joda-time-2.9.2.jar to DistributedCache through /tmp/11836216/tmp-113822416/joda-time-2.9.2.jar
2022-12-06 14:48:47,652 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Added jar file:/home/hadoopusr/pig-0.16.0/lib/joda-time-2.9.3.jar to DistributedCache through /tmp/11836216/tmp-113822416/joda-time-2.9.3.jar

```

```

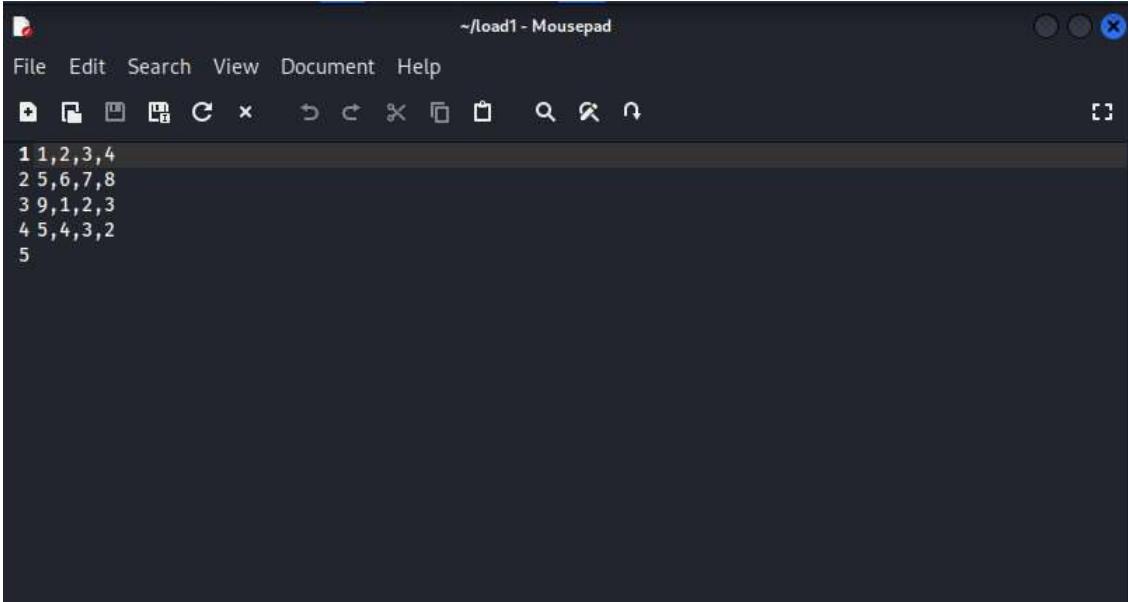
2022-12-06 14:48:53,271 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil
(2,8)
(7,8)
grunts>

```

### 3 Limit

The Apache Pig LIMIT operator is used to limit the number of output tuples. However, if you specify the limit of output tuples equal to or more than the number of tuples exists, all the tuples in the relation are returned.

- Create a text file in your local machine and insert the list of tuples.



Upload the text files on HDFS in the specific directory.

```
$ hdfs dfs -put plimit.txt /pigexample
```

```

(hadoopusr@kali)-[~/hadoop-2.9.1]
└─$ hadoop fs -put /home/hadoopusr/load1 /user/hadoopusr
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/common/lib/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
22/12/06 14:21:58  WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

```

- Load the file that contains the data.

```
$ pig
```

Rajeev Katara

```

(hadoopusr@kali)-[~/hadoop-2.9.1]
$ pig
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
22/12/06 15:01:20 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
22/12/06 15:01:20 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
22/12/06 15:01:20 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2022-12-06 15:01:20,315 [main] INFO org.apache.pig.Main - Apache Pig version 0.16.0 (r1746530) compiled Jun 01 2016, 23:10:49
2022-12-06 15:01:20,316 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoopusr/hadoop-2.9.1/pig_1670356880312.log
2022-12-06 15:01:20,378 [main] INFO org.apache.pig.impl.util.Utils - Default bootstrap file /home/hadoopusr/pigbootstrap not found
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/home/hadoopusr/hadoop-2.9.1/share/hadoop/com/b/hadoop-auth-2.9.1.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentication.util.KerberosUtil
WARNING: All illegal access operations will be denied in a future release
2022-12-06 15:01:21,323 [main] WARN org.apache.hadoop.util.NativeCodeLoader - Unable to load native-hadoop library for your platform... using built-in classes where applicable

```

o Load

the file that contains the data.

```
grunt> A = LOAD '/pigexample/plimit.txt' USING PigStorage(',') AS (a1:int,a2:int,a3:i);
```

```

2022-12-06 15:01:23,036 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> A = LOAD '/user/hadoopusr/load1' USING PigStorage(',') AS (a1:int,a2:int,a3:int);
2022-12-06 15:02:22,140 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> Dump A
2022-12-06 15:02:26,523 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2022-12-06 15:02:26,592 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-06 15:02:26,603 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-12-06 15:02:26,651 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2022-12-06 15:02:26,778 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (G1 Old Gen) of size 1048576000 to monitor. collectionUsageThreshold = 734003200, usageThreshold = 734003200
2022-12-06 15:02:26,901 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic
? false

```

Now, execute and verify the data. grunt> DUMP A;

```

2022-12-06 15:01:23,036 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> A = LOAD '/user/hadoopusr/load1' USING PigStorage(',') AS (a1:int,a2:int,a3:int);
2022-12-06 15:02:22,140 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
grunt> Dump A
2022-12-06 15:02:26,523 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: UNKNOWN
2022-12-06 15:02:26,592 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-06 15:02:26,603 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-12-06 15:02:26,651 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2022-12-06 15:02:26,778 [main] INFO org.apache.pig.impl.util.SpillableMemoryManager - Selected heap (G1 Old Gen) of size 1048576000 to monitor. collectionUsageThreshold = 734003200, usageThreshold = 734003200
2022-12-06 15:02:26,901 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic
? false

```

```

2022-12-06 15:02:26,000 [main] INFO org.apache.hadoop.hdfs.DFSClient - 
ready initialized
2022-12-06 15:02:36,008 [main] INFO org.apache.hadoop.hdfs.DFSClient - 
ready initialized
2022-12-06 15:02:36,019 [main] INFO org.apache.pig.backed.hdfs.DFSClient - 
ready initialized
2022-12-06 15:02:36,025 [main] INFO org.apache.hadoop.hdfs.DFSClient - 
ready initialized
2022-12-06 15:02:36,025 [main] WARN org.apache.pig.backed.hdfs.DFSClient - 
ready initialized
2022-12-06 15:02:36,037 [main] INFO org.apache.hadoop.hdfs.DFSClient - 
ready initialized
2022-12-06 15:02:36,037 [main] INFO org.apache.pig.backed.hdfs.DFSClient - 
ready initialized
(1,2,3)
(5,6,7)
(9,1,2)
(5,4,3)
grunt> ■

```

- o Let's return the first two tuples.

```
grunt> Result = LIMIT A 2;  grunt> DUMP Result;
```

```

(5,4,3)
grunt> Result = LIMIT A 2;
grunt> DUMP Result;
2022-12-06 15:04:45,671 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
2022-12-06 15:04:45,689 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-12-06 15:04:45,689 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-06 15:04:45,690 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalculator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, PartitionFilterOptimizer, PredicatePushdownOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter]}
2022-12-06 15:04:45,746 [main] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - File Output Committer Algorithm version is 1
2022-12-06 15:04:45,746 [main] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - FileOutputCommitter skip cleanup _temporary folders output directory:false, ignore cleanup failures: false
2022-12-06 15:04:45,755 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-12-06 15:04:45,782 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-06 15:04:45,783 [main] INFO org.apache.pig.builtin.PigStorage - Using PigTextInputFormat
2022-12-06 15:04:45,788 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-06 15:04:45,788 [main] INFO org.apache.pig.builtin.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
2022-12-06 15:04:46,235 [main] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt_0001_m_000001' to /localhost:54310/tmp/_temp-1393259920/tmp-1937660218/_temporary/0/task_0001_m_000001
2022-12-06 15:04:46,256 [main] WARN org.apache.pig.data.SchemaTupleBackend - SchemaTupleBackend has already been initialized
2022-12-06 15:04:46,268 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input files to process : 1
2022-12-06 15:04:46,268 [main] INFO org.apache.pig.backed.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(1,2,3)
(5,6,7)
grunt> ■

```

Rajeev Katara

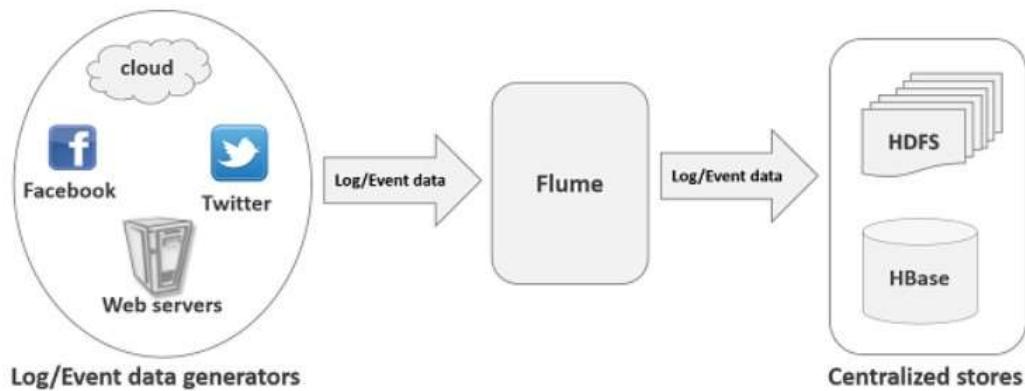
## Practical 7

**Aim:** Use Flume and transport the data from the various sources to a centralized data store.

### What is Apache Flume?

Apache Flume is a tool/service/data ingestion mechanism for collecting aggregating and transporting large amounts of streaming data such as log files, events (etc...) from various sources to a centralized data store.

Flume is a highly reliable, distributed, and configurable tool. It is principally designed to copy streaming data (log data) from various web servers to HDFS.



### Applications of Flume:

Assume an e-commerce web application wants to analyze the customer behavior from a particular region. To do so, they would need to move the available log data in to Hadoop for analysis. Here, Apache Flume comes to our rescue.

Flume is used to move the log data generated by application servers into HDFS at a higher speed.

### Features of Flume:

Some of the notable features of Flume are as follows –

- Flume ingests log data from multiple web servers into a centralized store (HDFS, HBase) efficiently.
- Using Flume, we can get the data from multiple servers immediately into Hadoop.
- Along with the log files, Flume is also used to import huge volumes of event data produced by social networking sites like Facebook and Twitter, and e-commerce websites like Amazon and Flipkart.
- Flume supports a large set of sources and destinations types.
- Flume supports multi-hop flows, fan-in fan-out flows, contextual routing, etc.
- Flume can be scaled horizontally.

### Advantages of Flume:

Here are the advantages of using Flume –

Rajeev Katara

- Using Apache Flume, we can store the data in to any of the centralized stores (HBase, HDFS).
- When the rate of incoming data exceeds the rate at which data can be written to the destination, Flume acts as a mediator between data producers and the centralized stores and provides a steady flow of data between them.
- Flume provides the feature of contextual routing.
- The transactions in Flume are channel-based where two transactions (one sender and one receiver) are maintained for each message. It guarantees reliable message delivery.
- Flume is reliable, fault tolerant, scalable, manageable, and customizable.

## Implementation:

### Pre-requisites:

Check the version of Java.

```
(mohammed@mohammed)-[~]
$ java -version
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
openjdk version "11.0.16" 2022-07-19
OpenJDK Runtime Environment (build 11.0.16+8-post-Debian-1)
OpenJDK 64-Bit Server VM (build 11.0.16+8-post-Debian-1, mixed mode, sharing)
```

Updating the linux distribution.

```
(mohammed@mohammed)-[~]
$ sudo apt-get install update
[sudo] password for mohammed:
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
E: Unable to locate package update
```

Updating the Java Development Kit (JDK).

```
(mohammed@mohammed)-[~]
$ sudo apt-get install default-jdk
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  default-jdk-headless libice-dev libpthread-stubs0-dev libsm-dev libx11-6 libx11-data
  libx11-dev libx11-xcb1 libxau-dev libxcb-damage0 libxcb-dri2-0 libxcb-dri3-0 libxcb-glx0
  libxcb-present0 libxcb-randr0 libxcb-render0 libxcb-shape0 libxcb-shm0 libxcb-sync1
  libxcb-xfixes0 libxcb-xinerama0 libxcb-xinput0 libxcb-xkb1 libxcb1 libxcb1-dev libxdmcp-dev
  libxt-dev openjdk-11-jdk openjdk-11-jdk-headless x11proto-dev xorg-sgml-doctools xtrans-dev
Suggested packages:
  libice-doc libsm-doc libx11-doc libxcb-doc libxt-doc openjdk-11-demo openjdk-11-source
  visualvm
The following NEW packages will be installed:
  default-jdk default-jdk-headless libice-dev libpthread-stubs0-dev libsm-dev libx11-dev
  libxau-dev libxcb1-dev libxdmcp-dev libxt-dev openjdk-11-jdk openjdk-11-jdk-headless
  x11proto-dev xorg-sgml-doctools xtrans-dev
The following packages will be upgraded:
  libx11-6 libx11-data libx11-xcb1 libxcb-damage0 libxcb-dri2-0 libxcb-dri3-0 libxcb-glx0
  libxcb-present0 libxcb-randr0 libxcb-render0 libxcb-shape0 libxcb-shm0 libxcb-sync1
  libxcb-xfixes0 libxcb-xinerama0 libxcb-xinput0 libxcb-xkb1 libxcb1
18 upgraded, 15 newly installed, 0 to remove and 1163 not upgraded.
Need to get 227 MB of archives.
After this operation, 239 MB of additional disk space will be used.
```

Then, download jdk1.8.0\_341 x64 Linux Distribution from Oracle.

```
(mohammed@Mohammed)-[~]
$ java -version
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
openjdk version "11.0.16" 2022-07-19
OpenJDK Runtime Environment (build 11.0.16+8-post-Debian-1)
OpenJDK 64-Bit Server VM (build 11.0.16+8-post-Debian-1, mixed mode, sharing)

(mohammed@Mohammed)-[~]
$ cd /usr/lib/jvm

(mohammed@Mohammed)-[/usr/lib/jvm]
$ sudo tar -xvzf ~/Downloads/jdk-8u341-linux-x64.tar.gz
[sudo] password for mohammed:
[jdk1.8.0_341/COPYRIGHT
[jdk1.8.0_341/LICENSE
[jdk1.8.0_341/README.html
[jdk1.8.0_341/THIRDPARTYLICENSEREADME.txt
[jdk1.8.0_341/bin/java-rmi.cgi
[jdk1.8.0_341/bin/appletviewer
[jdk1.8.0_341/bin/extcheck
[jdk1.8.0_341/bin/idlj
[jdk1.8.0_341/bin/jar
[jdk1.8.0_341/bin/jarsigner
[jdk1.8.0_341/bin/javac
[jdk1.8.0_341/bin/javadoc
[jdk1.8.0_341/bin/javah
[jdk1.8.0_341/bin/javap
[jdk1.8.0_341/bin/jdeps
[jdk1.8.0_341/bin/jconsole
[jdk1.8.0_341/bin/jdb
[jdk1.8.0_341/bin/jhat
[jdk1.8.0_341/bin/jinfo
[jdk1.8.0_341/bin/jmap
[jdk1.8.0_341/bin/jrungscript
[jdk1.8.0_341/bin/jps
[jdk1.8.0_341/bin/jjs
[jdk1.8.0_341/bin/jsadebugd
[jdk1.8.0_341/bin/jstack
[jdk1.8.0_341/bin/jstat
[jdk1.8.0_341/bin/jstatd
[jdk1.8.0_341/bin/keytool
```

Use the “update-alternatives” command to set the version of “java” and “javac”.

Make sure the versions of both java and javac are set to 1.8.0\_341.

```

└─(mohammed@Mohammed)-[/usr/lib/jvm]
$ cd jdk1.8.0_341

└─(mohammed@Mohammed)-[/usr/lib/jvm/jdk1.8.0_341]
$ sudo nano /etc/environment

└─(mohammed@Mohammed)-[/usr/lib/jvm/jdk1.8.0_341]
$ sudo nano /etc/environment

└─(mohammed@Mohammed)-[/usr/lib/jvm/jdk1.8.0_341]
$ sudo nano /etc/environment

└─(mohammed@Mohammed)-[/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --install "/usr/bin/java" "java" "/usr/lib/jvm/jdk1.8.0_341/bin/java" 0
update-alternatives: --install needs <link> <name> <path> <priority>

Use 'update-alternatives --help' for program usage information.

└─(mohammed@Mohammed)-[/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --install "/usr/bin/java" "java" "/usr/lib/jvm/jdk1.8.0_341/bin/java" 0

└─(mohammed@Mohammed)-[/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --install "/usr/bin/java" "java" "/usr/lib/jvm/jdk1.8.0_341/bin/java" 0
update-alternatives: using /usr/lib/jvm/jdk1.8.0_341/bin/java to provide /usr/bin/java (java) in
manual mode

└─(mohammed@Mohammed)-[/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --set java /usr/lib/jvm/jdk1.8.0_341/bin/java
update-alternatives: using /usr/lib/jvm/jdk1.8.0_341/bin/javac to provide /usr/bin/java (java) i
n manual mode

```

```

└─(mohammed@Mohammed)-[/usr/lib/jvm/jdk1.8.0_341]
$ sudo update-alternatives --config java
There are 3 choices for the alternative java (providing /usr/bin/java).

Selection    Path                      Priority    Status
-----      -----
0            /usr/lib/jvm/java-11-openjdk-amd64/bin/java    1111      auto mode
1            /usr/lib/jvm/java-11-openjdk-amd64/bin/java    1111      manual mode
2            /usr/lib/jvm/jdk1.8.0_341/bin/java             0          manual mode
* 3           /usr/lib/jvm/jdk1.8.0_341/bin/javac           0          manual mode

Press <enter> to keep the current choice[*], or type selection number: 2
update-alternatives: using /usr/lib/jvm/jdk1.8.0_341/bin/java to provide /usr/bin/java (java) in
manual mode

└─(mohammed@Mohammed)-[/usr/lib/jvm/jdk1.8.0_341]
$ java -version
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
java version "1.8.0_341"
Java(TM) SE Runtime Environment (build 1.8.0_341-b10)
Java HotSpot(TM) 64-Bit Server VM (build 25.341-b10, mixed mode)

```

```

[mohammed@Mohammed]~$ sudo update-alternatives --install "/usr/bin/javac" "javac" "/usr/lib/jvm/jdk1.8.0_341/bin/javac" 1
update-alternatives: using /usr/lib/jvm/jdk1.8.0_341/bin/javac to provide /usr/bin/javac (javac) in auto mode

[mohammed@Mohammed]~$ sudo update-alternatives --set javac /usr/lib/jvm/jdk1.8.0_341/bin/javac

[mohammed@Mohammed]~$ sudo update-alternatives --config javac
There is 1 choice for the alternative javac (providing /usr/bin/javac).

      Selection    Path          Priority   Status
* 1              /usr/lib/jvm/jdk1.8.0_341/bin/javac   1         manual mode
0              /usr/lib/jvm/jdk1.8.0_341/bin/javac   1         auto mode

Press <enter> to keep the current choice[*], or type selection number: 1

```

Now, check is java is perfectly installed.

```

[mohammed@Mohammed]~$ java -version
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
java version "1.8.0_341"
Java(TM) SE Runtime Environment (build 1.8.0_341-b10)
Java HotSpot(TM) 64-Bit Server VM (build 25.341-b10, mixed mode)

[mohammed@Mohammed]~$ javac -version
Picked up _JAVA_OPTIONS: -Dawt.useSystemAAFontSettings=on -Dswing.aatext=true
javac 1.8.0_341

```

Create a separate user on linux to install Hadoop. We create a user “hadoopusr” in the usergroup “hadoop”.

```

[mohammed@Mohammed]~$ sudo addgroup hadoop
[sudo] password for mohammed:
Adding group `hadoop' (GID 1001) ...
Done.

[mohammed@Mohammed]~$ sudo adduser --ingroup hadoop hadoopusr
Adding user `hadoopusr' ...
Adding new user `hadoopusr' (1001) with group `hadoop' ...
Creating home directory `/home/hadoopusr' ...
Copying files from `/etc/skel' ...
New password:
Retype new password:
passwd: password updated successfully
Changing the user information for hadoopusr
Enter the new value, or press ENTER for the default
  Full Name []: Apache Flume Project
  Room Number []:
  Work Phone []:
  Home Phone []:
  Other []:
Is the information correct? [Y/n] Y

[mohammed@Mohammed]~$ sudo adduser hadoopusr sudo
Adding user `hadoopusr' to group `sudo' ...
Done.

[mohammed@Mohammed]~$ 

```

Now we will install the SSH Server.

```
(mohammed@Mohammed)~]$ sudo apt-get install openssh-server
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  libc-bin libc-dev-bin libc-l10n libc6 libc6-dev libc6-i386 libssl3 locales openssh-client
  openssh-sftp-server openssl runit-helper
Suggested packages:
  glibc-doc libnss-nis libnss-nisplus manpages-dev keychain libpam-ssh monkeysphere
  ssh-askpass molly-guard ufw
Recommended packages:
  manpages-dev libc-devtools
The following packages will be upgraded:
  libc-bin libc-dev-bin libc-l10n libc6 libc6-dev libc6-i386 libssl3 locales openssh-client
  openssh-server openssh-sftp-server openssl runit-helper
13 upgraded, 0 newly installed, 0 to remove and 1168 not upgraded.
Need to get 17.2 MB of archives.
After this operation, 4,006 kB disk space will be freed.
Do you want to continue? [Y/n] Y
```

Now we will switch users from “kali” to “hadoopusr”.

```
(mohammed@Mohammed)~]$ su - hadoopusr
Password:
(hadoopusr@Mohammed)~]$
```

Then we will create and save a SSH key.

```
(hadoopusr@Mohammed)~]$ ssh-keygen -t rsa -P ""
Generating public/private rsa key pair.
Enter file in which to save the key (/home/hadoopusr/.ssh/id_rsa):
Created directory '/home/hadoopusr/.ssh'.
Your identification has been saved in /home/hadoopusr/.ssh/id_rsa
Your public key has been saved in /home/hadoopusr/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:DVH/6ZIUp7sTlWktJ4xpm0+/ygv/E5j0upI4p0j5t+A hadoopusr@Mohammed
The key's randomart image is:
+---[RSA 3072]---+
| ...
| . .
| . o ...
| o o=+.
| S .+=o
| . +B+o
| ..o . o.=+..
| .o..+ o o=+.
| .o.E..o .o+B=+o
+---[SHA256]---+
(hadoopusr@Mohammed)~]$ cat $HOME/.ssh/id_rsa.pub >> $HOME/.ssh/authorized_keys
-bash: /home/hadoopusr/: Is a directory
```

Now, we will start a SSH server on localhost.

```

└─(hadoopusr@Mohammed)-[~]
└─$ sudo service ssh status
[sudo] password for hadoopusr:
● ssh.service - OpenBSD Secure Shell server
  Loaded: loaded (/lib/systemd/system/ssh.service; disabled; vendor preset: disabled)
  Active: inactive (dead)
    Docs: man:sshd(8)
          man:sshd_config(5)

└─(hadoopusr@Mohammed)-[~]
└─$ sudo service ssh start

└─(hadoopusr@Mohammed)-[~]
└─$ sudo service ssh status
● ssh.service - OpenBSD Secure Shell server
  Loaded: loaded (/lib/systemd/system/ssh.service; disabled; vendor preset: disabled)
  Active: active (running) since Thu 2022-12-01 21:33:00 IST; 8s ago
    Docs: man:sshd(8)
          man:sshd_config(5)
  Process: 4589 ExecStartPre=/usr/sbin/sshd -t (code=exited, status=0/SUCCESS)
  Main PID: 4590 (sshd)
    Tasks: 1 (limit: 2214)
   Memory: 2.8M
      CPU: 71ms
     CGroup: /system.slice/ssh.service
             └─4590 "sshd: /usr/sbin/sshd -D [listener] 0 of 10-100 startups"

Dec 01 21:33:00 Mohammed systemd[1]: Starting OpenBSD Secure Shell server ...
Dec 01 21:33:00 Mohammed sshd[4590]: Server listening on 0.0.0.0 port 22.
Dec 01 21:33:00 Mohammed sshd[4590]: Server listening on :: port 22.
Dec 01 21:33:00 Mohammed systemd[1]: Started OpenBSD Secure Shell server.

└─(hadoopusr@Mohammed)-[~]
└─$ ssh localhost
The authenticity of host 'localhost (::1)' can't be established.
ED25519 key fingerprint is SHA256:UkJN1l6R7QcR05P9w67wvSco3jBtcHPhjJ8EbD2B30g.
This key is not known by any other names
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
Warning: Permanently added 'localhost' (ED25519) to the list of known hosts.
hadoopusr@localhost's password:
Linux Mohammed 5.18.0-kali5-amd64 #1 SMP PREEMPT_DYNAMIC Debian 5.18.5-1kali6 (2022-07-07) x8
4

The programs included with the Kali GNU/Linux system are free software;
the exact distribution terms for each program are described in the
individual files in /usr/share/doc/*copyright.

Kali GNU/Linux comes with ABSOLUTELY NO WARRANTY, to the extent
permitted by applicable law.

```

Now we will exit the “hadoopusr”.

```

└─(hadoopusr@Mohammed)-[~]
└─$ exit
logout
Connection to localhost closed.

└─(hadoopusr@Mohammed)-[~]
└─$ █

```

Now, we will download and install Apache Hadoop 2.9.0.

Index of /dist/hadoop/common/ +

← → ⌂ ⌂ https://archive.apache.org/dist/hadoop/common/hadoop-2.9.0/ ⌂ ⌂

Kali Linux Kali Tools Kali Docs Kali Forums Kali NetHunter Exploit-DB Google Hacking DB OffSec

## Index of /dist/hadoop/common/hadoop-2.9.0

Name	Last modified	Size	Description
Parent Directory		-	
<a href="#">hadoop-2.9.0-src.tar.gz</a>	2017-11-17 23:08	37M	
<a href="#">hadoop-2.9.0-src.tar.gz.asc</a>	2017-11-17 23:08	819	
<a href="#">hadoop-2.9.0-src.tar.gz.md5</a>	2017-11-17 23:08	162	
<a href="#">hadoop-2.9.0-src.tar.gz.mds</a>	2017-11-17 23:08	1.0K	
<a href="#">hadoop-2.9.0-src.tar.gz.sha256</a>	2018-03-13 20:32	90	
<a href="#">hadoop-2.9.0.tar.gz</a>	2017-11-17 23:10	350M	
<a href="#">hadoop-2.9.0.tar.gz.asc</a>	2017-11-17 23:08	819	
<a href="#">hadoop-2.9.0.tar.gz.md5</a>	2017-11-17 23:08	154	
<a href="#">hadoop-2.9.0.tar.gz.mds</a>	2017-11-17 23:08	1.0K	
<a href="#">hadoop-2.9.0.tar.gz.sha256</a>	2018-03-13 20:32	86	

```
(hadoopusr@Mohammed)@[~/Desktop]
$ cd ~/Downloads

(hadoopusr@Mohammed)@[~/Downloads]
$ ls
hadoop-2.9.0.tar.gz jdk-8u341-linux-x64.tar.gz

(hadoopusr@Mohammed)@[~/Downloads]
$ sudo tar -xvzf hadoop-2.9.0.tar.gz
[sudo] password for hadoopusr:
hadoop-2.9.0/
hadoop-2.9.0/include/
hadoop-2.9.0/include/Pipes.hh
hadoop-2.9.0/include/SerialUtils.hh
hadoop-2.9.0/include/hdfs.h
hadoop-2.9.0/include/StringUtils.hh
hadoop-2.9.0/include/TemplateFactory.hh
hadoop-2.9.0/NOTICE.txt
hadoop-2.9.0/lib/
hadoop-2.9.0/lib/native/
hadoop-2.9.0/lib/native/libhdfs.a
hadoop-2.9.0/lib/native/libhadoop.so
hadoop-2.9.0/lib/native/libhadoop.so.1.0.0
hadoop-2.9.0/lib/native/libhdfs.so.0.0.0
hadoop-2.9.0/lib/native/libhadoop.a
hadoop-2.9.0/lib/native/libhdfs.so
hadoop-2.9.0/lib/native/examples/
hadoop-2.9.0/lib/native/examples/pipes-sort
hadoop-2.9.0/lib/native/examples/wordcount-part
hadoop-2.9.0/lib/native/examples/wordcount-simple
hadoop-2.9.0/lib/native/examples/wordcount-nopipe
hadoop-2.9.0/lib/native/libhadooputils.a
hadoop-2.9.0/lib/native/libhadooppipes.a
hadoop-2.9.0/LICENSE.txt
hadoop-2.9.0/libexec/
hadoop-2.9.0/libexec/kms-config.sh
hadoop-2.9.0/libexec/mapred-config.cmd
hadoop-2.9.0/libexec/mapred-config.sh
hadoop-2.9.0/libexec/hadoop-config.cmd
hadoop-2.9.0/libexec/hadoop-config.sh
```

Now we will move the file to the “hadoopusr” and transfer ownership to the “hadoopusr”.

```

└─(hadoopusr@Mohammed)-[~/home/mohammed/Downloads]
$ ls
hadoop-2.9.0  hadoop-2.9.0.tar.gz  jdk-8u341-linux-x64.tar.gz

└─(hadoopusr@Mohammed)-[~/home/mohammed/Downloads]
$ sudo mv hadoop-2.9.0 /usr/local/hadoop

└─(hadoopusr@Mohammed)-[~/home/mohammed/Downloads]
$ sudo chown -R hadoopusr /usr/local

└─(hadoopusr@Mohammed)-[~/home/mohammed/Downloads]
$ █

```

Now make configurations to the “bashrc” file using nano.

```

└─(hadoopusr@Mohammed)-[~/home/mohammed/Downloads]
$ sudo nano ~/.bashrc

GNU nano 6.3                               /home/hadoopusr/.bashrc *

# some more ls aliases
alias ll='ls -l'
alias la='ls -A'
alias l='ls -CF'

# Alias definitions.
# You may want to put all your additions into a separate file like
# ~/.bash_aliases, instead of adding them here directly.
# See /usr/share/doc/bash-doc/examples in the bash-doc package.

if [ -f ~/.bash_aliases ]; then
    . ~/.bash_aliases
fi

# enable programmable completion features (you don't need to enable
# this, if it's already enabled in /etc/bash.bashrc and /etc/profile
# sources /etc/bash.bashrc).
if ! shopt -oq posix; then
    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
export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_341
export JAVA_PATH=$PATH:$JAVA_HOME/bin
export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$HADOOP_HOME/bin
export PATH=$PATH:$HADOOP_HOME/sbin
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=$HADOOP_HOME/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/native"

^G Help          ^O Write Out      ^W Where Is      ^K Cut          ^T Execute      ^C Location
^X Exit          ^R Read File     ^\ Replace       ^U Paste        ^J Justify      ^/ Go To Line

```

Then we will run “source ~/.bashrc” to enable the changes.

```

└─(hadoopusr@Mohammed)-[~/home/mohammed/Downloads]
$ source ~/.bashrc

└─(hadoopusr@Mohammed)-[~/home/mohammed/Downloads]
$ █

```

Next, we will make some changes to the hadoop environment file.

We will run the command “sudo nano /usr/local/hadoop/etc/hadoop/hadoop-env.sh”.

```
(hadoopusr@Mohammed)-[~/home/mohammed/Downloads]
$ sudo nano /usr/local/hadoop/etc/hadoop/hadoop-env.sh

(hadoopusr@Mohammed)-[~/home/mohammed/Downloads]
$
```

Here we will set the path of JAVA\_HOME to “JAVA\_HOME=/usr/lib/jvm/jdk1.8.0\_341”.

```
GNU nano 6.3          /usr/local/hadoop/etc/hadoop/hadoop-env.sh *
# 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.

# Set Hadoop-specific environment variables here.

# The only required environment variable is JAVA_HOME. All others are
# optional. When running a distributed configuration it is best to
# set JAVA_HOME in this file, so that it is correctly defined on
# remote nodes.

# The java implementation to use.
#export JAVA_HOME=${JAVA_HOME}
export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_341

# The jsvc implementation to use. Jsvc is required to run secure datanodes
# that bind to privileged ports to provide authentication of data transfer
# protocol. Jsvc is not required if SASL is configured for authentication of
# data transfer protocol using non-privileged ports.
#export JSVC_HOME=${JSVC_HOME}

export HADOOP_CONF_DIR=${HADOOP_CONF_DIR:-"/etc/hadoop"}

# Extra Java CLASSPATH elements. Automatically insert capacity-scheduler.
for f in $HADOOP_HOME/contrib/capacity-scheduler/*.jar; do
  if [ "$HADOOP_CLASSPATH" ]; then
    export HADOOP_CLASSPATH=$HADOOP_CLASSPATH:$f
  else
    export HADOOP_CLASSPATH=$f
  fi
done

# The maximum amount of heap to use, in MB. Default is 1000.
#export HADOOP_HEAPSIZE=
#export HADOOP_NAMENODE_INIT_HEAPSIZE=""

# Enable extra debugging of Hadoop's JAAS binding, used to set up

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

Now we will make the following changes to the configuration of the core-site.xml file.

```
<property>
  <name>fs.default.name</name>
  <value>hdfs://localhost:9000</value>
</property>
```

```
(hadoopusr@Mohammed)-[/usr/local/hadoop/etc/hadoop]
$ sudo nano core-site.xml
[sudo] password for hadoopusr:
```

```
GNU nano 6.3                                     core-site.xml *
```

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
    Licensed under the Apache License, Version 2.0 (the "License");
    you may not use this file except in compliance with the License.
    You may obtain a copy of the License at

        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>fs.default.name</name>
    <value>hdfs://localhost:9000</value>
</property>
</configuration>
```

File Edit View Insert Cell Help

^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location  
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line

Now we will make the following changes to the configuration of the hdfs-site.xml file.

```
<property>
    <name>dfs.replication</name>
    <value>1</value>
</property>
<property>
    <name>dfs.name.name.dir</name>
    <value>file:/usr/local/hadoop_tmp/hdfs/namenode</value>
</property>
<property>
    <name>dfs.data.data.dir</name>
    <value>file:/usr/local/hadoop_tmp/hdfs/datanode</value>
</property>
```

```
(hadoopusr@Mohammed)-[/usr/local/hadoop/etc/hadoop]
$ sudo nano hdfs-site.xml
```

```

GNU nano 6.3                               hdfs-site.xml *
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

    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>dfs.replication</name>
    <value>1</value>
</property>
<property>
    <name>dfs.name.name.dir</name>
    <value>file:/usr/local/hadoop_tmp/hdfs/namenode</value>
</property>
<property>
    <name>dfs.data.data.dir</name>
    <value>file:/usr/local/hadoop_tmp/hdfs/datanode</value>
</property>
</configuration>

^G Help          ^O Write Out      ^W Where Is      ^K Cut          ^T Execute      ^C Location
^X Exit          ^R Read File       ^\ Replace       ^U Paste        ^J Justify      ^/ Go To Line

```

Now we will make the following changes to the configuration of the yarn-site.xml file.

```

<property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
</property>

```

```

└─(hadoopusr@Mohammed)─[/usr/local/hadoop/etc/hadoop]
$ sudo nano yarn-site.xml
[sudo] password for hadoopusr:

```

```
GNU nano 6.3                               yarn-site.xml *
<?xml version="1.0"?>
<!—
    Licensed under the Apache License, Version 2.0 (the "License");
    you may not use this file except in compliance with the License.
    You may obtain a copy of the License at

        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.
→
<configuration>
<property>
    <name>yarn.nodemanager.aux-services</name>
    <value>mapreduce_shuffle</value>
</property>

<!— Site specific YARN configuration properties —>

</configuration>■
```

File Edit View Insert Cell Help

^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location  
^X Exit ^R Read File ^N Replace ^U Paste ^J Justify ^/ Go To Line

Now we will make the following changes to the configuration of the mapred-site.xml file.

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

```
(hadoopusr@Mohammed)-[/usr/local/hadoop/etc/hadoop]
$ sudo nano mapred-site.xml
```

```

GNU nano 6.3                               mapred-site.xml *
<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
Licensed under the Apache License, Version 2.0 (the "License");
you may not use this file except in compliance with the License.
You may obtain a copy of the License at

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>

```

(1) 1000x1000px image of a terminal window showing the nano editor with the mapred-site.xml configuration file. The file contains XML code for Hadoop's MapReduce framework, specifically setting the framework to YARN. The terminal has a dark background with light-colored text. A status bar at the bottom shows various keyboard shortcuts.

Next, we will create the following directories and transfer the ownership to the “hadoopusr”. These directories are where hadoop will store data about the namenode and the datanode.

```

└─(hadoopusr@Mohammed)-[ /usr/local/hadoop/etc/hadoop ]
$ cd /home/mohammed/Downloads
└─(hadoopusr@Mohammed)-[ /home/mohammed/Downloads ]
$ sudo mkdir -p /usr/local/hadoop_space
└─(hadoopusr@Mohammed)-[ /home/mohammed/Downloads ]
$ sudo mkdir -p /usr/local/hadoop_space/hdfs/namenode
└─(hadoopusr@Mohammed)-[ /home/mohammed/Downloads ]
$ sudo mkdir -p /usr/local/hadoop_space/hdfs/datanode
└─(hadoopusr@Mohammed)-[ /home/mohammed/Downloads ]
$ sudo chown -R hadoopusr /usr/local/hadoop_space
└─(hadoopusr@Mohammed)-[ /home/mohammed/Downloads ]
$ █

```

Then we format the namenode in the HDFS file structure.

```

[hadoopusr@Mohammed]~[~/home/mohammed/Downloads]
$ cd

[hadoopusr@Mohammed]~[~]
$ hdfs namenode -format
22/12/02 17:51:54 INFO namenode.NameNode: STARTUP_MSG:
*****STARTUP_MSG: Starting NameNode
STARTUP_MSG: host = Mohammed/127.0.1.1
STARTUP_MSG: args = [-format]
STARTUP_MSG: version = 2.9.0
STARTUP_MSG: classpath = /usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/hadoop/common/lib/stax-api-1.0-2.jar:/usr/local/hadoop/share/hadoop/common/lib/api-asn1-api-1.0.0-M20.jar:/usr/local/hadoop/share/hadoop/common/lib/avro-1.7.7.jar:/usr/local/hadoop/share/hadoop/common/lib/json-jaxrs-1.9.13.jar:/usr/local/hadoop/share/hadoop/common/lib/snappy-java-1.0.5.jar:/usr/local/hadoop/share/hadoop/common/lib/json-smart-1.1.1.jar:/usr/local/hadoop/share/hadoop/common/lib/protobuf-java-2.5.0.jar:/usr/local/hadoop/share/hadoop/common/lib/apacheds-i18n-2.0.0-M15.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-lang3-3.4.jar:/usr/local/hadoop/share/hadoop/common/lib/jaxb-api-2.2.2.jar:/usr/local/hadoop/share/hadoop/common/lib/servlet-api-2.5.jar:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar:/usr/local/hadoop/share/hadoop/common/lib/jackson-mapper-asl-1.9.13.jar:/usr/local/hadoop/share/hadoop/common/lib/hadoop-annotations-2.9.0.jar:/usr/local/hadoop/share/hadoop/common/lib/zookeeper-3.4.6.jar:/usr/local/hadoop/share/hadoop/common/lib/jetty-util-6.1.26.jar:/usr/local/hadoop/share/hadoop/common/lib/hamcrest-core-1.3.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-lang-2.6.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-configuration-1.6.jar:/usr/local/hadoop/share/hadoop/common/lib/gson-2.2.4.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-beanutils-core-1.8.0.jar:/usr/local/hadoop/share/hadoop/common/lib/jettison-1.1.jar:/usr/local/hadoop/share/hadoop/common/lib/activation-1.1.jar:/usr/local/hadoop/share/hadoop/common/lib/netty-3.6.2.Final.jar:/usr/local/hadoop/share/hadoop/common/lib/xz-1.0.jar:/usr/local/hadoop/share/hadoop/common/lib/guava-11.0.2.jar:/usr/local/hadoop/share/hadoop/common/lib/curator-framework-2.7.1.jar:/usr/local/hadoop/share/hadoop/common/lib/httpcore-4.4.4.jar:/usr/local/hadoop/share/hadoop/common/lib/jersey-json-1.9.jar:/usr/local/hadoop/share/hadoop/common/lib/paranamer-2.3.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-logging-1.1.3.jar:/usr/local/hadoop/share/hadoop/common/lib/commons-math3-3.1.1.jar:/usr/local/hadoop/share/hadoop/common/lib/slf4j-api-1.7.25.jar:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.9.0.jar:/usr/local/hadoop/share/hadoop/common/lib/jsr305-3.0.0.jar:/usr/local/hadoop/share/hadoop/common/lib/jetty-sslengine-6.1.26.jar:/usr/local/hadoop/share/hadoop/common/lib/xmlenc-0.52.jar:/usr/local

```

Next, we will start the Distributed file system.

```

[hadoopusr@Mohammed]~[~]
$ start-dfs.sh
22/12/02 17:56:24 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Starting namenodes on [localhost]
hadoopusr@localhost's password:
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-hadoopusr-namenode-Mohammed.out
hadoopusr@localhost's password:
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-hadoopusr-datanode-Mohammed.out
Starting secondary namenodes [0.0.0.0]
The authenticity of host '0.0.0.0 (0.0.0.0)' can't be established.
ED25519 key fingerprint is SHA256:UkJN1l6R7QcR05P9w67wvSc03jBtCHPhjJ8EbD2B3Og.
This host key is known by the following other names/addresses:
    ~/.ssh/known_hosts:1: [hashed name]
Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
0.0.0.0: Warning: Permanently added '0.0.0.0' (ED25519) to the list of known hosts.
hadoopusr@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-hadoopusr-secondarynamenode-Mohammed.out
22/12/02 17:57:47 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

```

Then the YARN system.

```
(hadoopusr@Mohammed)-[~]
$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-hadoopusr-resourcemanager-Moham
med.out
hadoopusr@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-hadoopusr-nodemanager-Mo
hammed.out
```

Check if all the systems are running.

```
(hadoopusr@Mohammed)-[~]
$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
22/12/02 18:00:12 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your plat
form ... using builtin-java classes where applicable
Starting namenodes on [localhost]
hadoopusr@localhost's password:
localhost: namenode running as process 22065. Stop it first.
hadoopusr@localhost's password:
localhost: datanode running as process 22192. Stop it first.
Starting secondary namenodes [0.0.0.0]
hadoopusr@0.0.0.0's password:
0.0.0.0: secondarynamenode running as process 22581. Stop it first.
22/12/02 18:00:48 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your plat
form ... using builtin-java classes where applicable
starting yarn daemons
resourcemanager running as process 22984. Stop it first.
hadoopusr@localhost's password:
localhost: nodemanager running as process 23170. Stop it first.
```

Here we can see that hadoop has been successfully installed and activated We have to go the url <http://localhost:50070> to access the Hadoop Server page.

<b>Started:</b>	Fri Dec 02 17:56:43 +0530 2022
<b>Version:</b>	2.9.0, r756ebc8394e473ac25feac05fa493f6d612e6c50
<b>Compiled:</b>	Tue Nov 14 04:45:00 +0530 2017 by arsuresh from branch-2.9.0
<b>Cluster ID:</b>	CID-8a6838e1-ff11-4c7f-9efb-a4473c2272e6
<b>Block Pool ID:</b>	BP-1774126645-127.0.1.1-1669983717900

<b>Configured Capacity:</b>	57.32 GB
<b>DFS Used:</b>	24 KB (0%)
<b>Non DFS Used:</b>	13.96 GB

We can also see the information of the cluster.

The screenshot shows the Hadoop 'All Applications' interface. On the left, there's a sidebar with navigation links like 'About', 'Nodes', 'Node Labels', 'Applications', 'Scheduler', and 'Tools'. The main area has tabs for 'Cluster Metrics', 'Cluster Nodes Metrics', and 'Scheduler Metrics'. Under 'Cluster Metrics', there's a table with columns: Apps Submitted, Apps Pending, Apps Running, Apps Completed, Containers Running, Memory Used, Memory Total, Memory Reserved, VCores Used, VCores Total, and VCores Reserved. All values are 0. Under 'Scheduler Metrics', it shows the Capacity Scheduler with [MEMORY] as the Scheduling Resource Type, and allocation ranges <memory:1024, vCores:1> to <memory:8192, vCores:4>. A table below lists application details with 0 entries.

Next, we move on to the process of downloading and installing Apache Flume. I downloaded Apache Flume version 1.11.0 and performed the following:

```
(mohammed@mohammed)-[~]
$ cd /home/mohammed/Downloads/
(mohammed@mohammed)-[~/Downloads]
$ tar -xvf apache-flume-1.11.0-bin.tar.gz
apache-flume-1.11.0-bin/LICENSE
apache-flume-1.11.0-bin/NOTICE
apache-flume-1.11.0-bin/bin/
apache-flume-1.11.0-bin/conf/
apache-flume-1.11.0-bin/DEVNOTES
apache-flume-1.11.0-bin/bin/flume-ng.cmd
apache-flume-1.11.0-bin/bin/flume-ng
apache-flume-1.11.0-bin/bin/flume-ng.ps1
apache-flume-1.11.0-bin/CHANGELOG
apache-flume-1.11.0-bin/RELEASE-NOTES
apache-flume-1.11.0-bin/README.md
apache-flume-1.11.0-bin/doap_Flume.rdf
apache-flume-1.11.0-bin/conf/flume-env.ps1.template
apache-flume-1.11.0-bin/conf/log4j2.xml
apache-flume-1.11.0-bin/conf/flume-conf.properties.template
apache-flume-1.11.0-bin/conf/flume-env.sh.template
apache-flume-1.11.0-bin/docs/
apache-flume-1.11.0-bin/docs/css/
apache-flume-1.11.0-bin/docs/images/
apache-flume-1.11.0-bin/docs/images/logos/
apache-flume-1.11.0-bin/docs/apidocs/
apache-flume-1.11.0-bin/docs/apidocs/org/
apache-flume-1.11.0-bin/docs/apidocs/org/apache/
apache-flume-1.11.0-bin/docs/apidocs/org/apache/flume/
apache-flume-1.11.0-bin/docs/apidocs/org/apache/flume/class-use/
apache-flume-1.11.0-bin/docs/apidocs/org/apache/flume/clients/
apache-flume-1.11.0-bin/docs/apidocs/org/apache/flume/clients/log4jappender/
apache-flume-1.11.0-bin/docs/apidocs/org/apache/flume/clients/log4jappender/class-use/
apache-flume-1.11.0-bin/docs/apidocs/org/apache/flume/tools/
apache-flume-1.11.0-bin/docs/apidocs/org/apache/flume/tools/class-use/
apache-flume-1.11.0-bin/docs/apidocs/org/apache/flume/util/
```

Then, we make changes to the `~/.bashrc` file.

```
(mohammed@Mohammed)@[~]
└─$ su - hadoopusr
Password:
(hadoopusr@Mohammed)@[~]
└─$ sudo nano ~/.bashrc
[sudo] password for hadoopusr:
```

```
fi
export JAVA_HOME=/usr/lib/jvm/jdk1.8.0_341
export JAVA_PATH=$PATH:$JAVA_HOME/bin
export HADOOP_HOME=/usr/local/hadoop
export PATH=$PATH:$HADOOP_HOME/bin
export HADOOP_HOME=$HADOOP_HOME/sbin
export HADOOP_MAPRED_HOME=$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/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/native"

export FLUME_HOME=/home/mohammed/Downloads/apache-flume-1.11.0-bin/
export PATH=$PATH:$FLUME_HOME/bin

^G Help      ^O Write Out    ^W Where Is    ^K Cut        ^T Execute   ^C Location
^X Exit      ^R Read File    ^N Replace     ^U Paste      ^J Justify    ^/ Go To Line
```

Now, we check the version of Flume and hence know it is perfectly installed.

```
(hadoopusr@Mohammed)@[~]
└─$ source .bashrc
(hadoopusr@Mohammed)@[~]
└─$ flume-ng version
Flume 1.11.0
Source code repository: https://git.apache.org/repos/asf/flume.git
Revision: 1a15927e594fd0d05a59d804b90a9c31ec93f5e1
Compiled by rgoers on Sun Oct 16 14:44:15 MST 2022
From source with checksum bbbca682177262aac3a89defde369a37

(hadoopusr@Mohammed)@[~]
```

## Now, we will execute the steps for Netcat Source using Flume

We have to configure the source, the channel, and the sink using the configuration file in the conf folder. Here, we use a NetCat Source, Memory channel, and a logger sink.

For this, we create a file named “netcat.conf” in the conf folder of our Flume.

```
mohammed@mohammed-VirtualBox: ~ $ cd ~/Downloads/apache-flume-1.9.0-bin/conf
mohammed@mohammed-VirtualBox: ~/Downloads/apache-flume-1.9.0-bin/conf $ sudo nano
netcat.conf
```

Then provide the following configuration in the netcat.conf file.

```
# Naming the components on the current agent
NetcatAgent.sources = Netcat
NetcatAgent.channels = MemChannel
```

```

NetcatAgent.sinks = LoggerSink

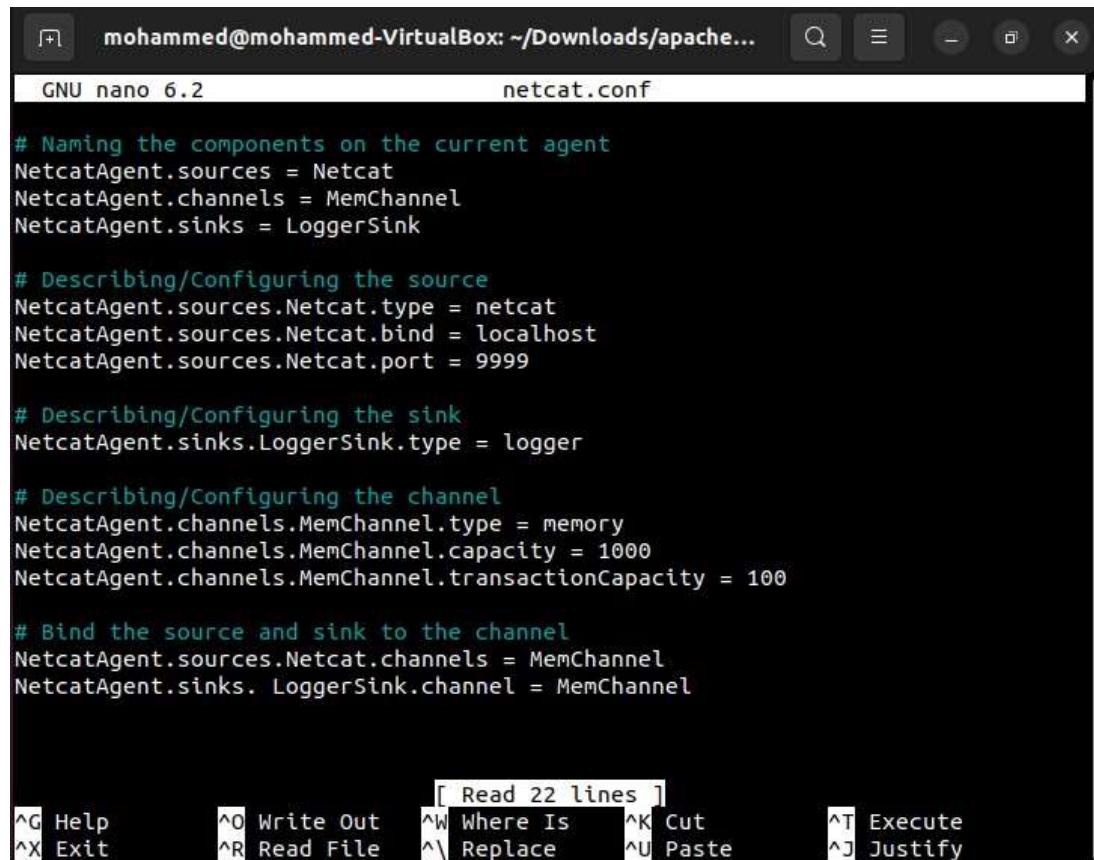
# Describing/Configuring the source
NetcatAgent.sources.Netcat.type = netcat
NetcatAgent.sources.Netcat.bind = localhost
NetcatAgent.sources.Netcat.port = 56565

# Describing/Configuring the sink
NetcatAgent.sinks.LoggerSink.type = logger

# Describing/Configuring the channel
NetcatAgent.channels.MemChannel.type = memory
NetcatAgent.channels.MemChannel.capacity = 1000
NetcatAgent.channels.MemChannel.transactionCapacity = 100

# Bind the source and sink to the channel
NetcatAgent.sources.Netcat.channels = MemChannel
NetcatAgent.sinks.LoggerSink.channel = MemChannel

```



```

mohammed@mohammed-VirtualBox: ~/Downloads/apache...
GNU nano 6.2          netcat.conf

# Naming the components on the current agent
NetcatAgent.sources = Netcat
NetcatAgent.channels = MemChannel
NetcatAgent.sinks = LoggerSink

# Describing/Configuring the source
NetcatAgent.sources.Netcat.type = netcat
NetcatAgent.sources.Netcat.bind = localhost
NetcatAgent.sources.Netcat.port = 9999

# Describing/Configuring the sink
NetcatAgent.sinks.LoggerSink.type = logger

# Describing/Configuring the channel
NetcatAgent.channels.MemChannel.type = memory
NetcatAgent.channels.MemChannel.capacity = 1000
NetcatAgent.channels.MemChannel.transactionCapacity = 100

# Bind the source and sink to the channel
NetcatAgent.sources.Netcat.channels = MemChannel
NetcatAgent.sinks.LoggerSink.channel = MemChannel

[ Read 22 lines ]
^G Help      ^O Write Out   ^W Where Is   ^K Cut       ^T Execute
^X Exit      ^R Read File    ^\ Replace    ^U Paste     ^J Justify

```

Browse through the Flume home directory and execute the application as shown below.

```
mohammed@mohammed-VirtualBox:~/Downloads/apache-flume-1.9.0-bin/conf$ cd $FLUME_HOME
mohammed@mohammed-VirtualBox:~/Downloads/apache-flume-1.9.0-bin$ flume-ng agent -n NetcatAgent -f /home/mohammed/Downloads/apache-flume-1.9.0-bin/conf/netcat.conf
Warning: No configuration directory set! Use --conf <dir> to override.
Info: Including Hadoop libraries found via (/home/mohammed/hadoop-3.2.3//bin/hadoop) for HDFS access
/home/mohammed/hadoop-3.2.3//libexec/hadoop-functions.sh: line 2366: HADOOP_ORG.APACHE.FLUME.TOOLS.GETJAVAPROPERTY_USER: invalid variable name
/home/mohammed/hadoop-3.2.3//libexec/hadoop-functions.sh: line 2461: HADOOP_ORG.APACHE.FLUME.TOOLS.GETJAVAPROPERTY_OPTS: invalid variable name
Info: Including Hive libraries found via () for Hive access
+ exec /usr/lib/jvm/java-8-openjdk-amd64/bin/java -Xmx20m -cp '/home/mohammed/Downloads/apache-flume-1.9.0-bin/lib/*:/home/mohammed/hadoop-3.2.3//etc/hadoop:/home/mohammed/hadoop-3.2.3//share/hadoop/common/lib/*:/home/mohammed/hadoop-3.2.3//share/hadoop/common/*:/home/mohammed/hadoop-3.2.3//share/hadoop/hdfs:/home/mohammed/hadoop-3.2.3//share/hadoop/hdfs/lib/*:/home/mohammed/hadoop-3.2.3//share/hadoop/mapreduce/lib/*:/home/mohammed/hadoop-3.2.3//share/hadoop/mapreduce/*:/home/mohammed/hadoop-3.2.3//share/hadoop/yarn:/home/mohammed/hadoop-3.2.3//share/hadoop/yarn/lib/*:/home/mohammed/hadoop-3.2.3//share/hadoop/yarn/*:/lib/*' -Djava.library.path=:/home/mohammed/hadoop-3.2.3//lib/native org.apache.flume.node.Application -n NetcatAgent -f /home/mohammed/Downloads/apache-flume-1.9.0-bin/conf/netcat.conf
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/mohammed/Downloads/apache-flume-1.9.0-b
```

The following snippet at the end of this execution shows us that the socket connection is successful.

```
2022-12-07 17:03:04,444 INFO instrumentation.MonitoredCounterGroup: Component type: CHANNEL, name: MemChannel started
2022-12-07 17:03:04,445 INFO node.Application: Starting Source Netcat
2022-12-07 17:03:04,483 INFO source.NetcatSource: Source starting
2022-12-07 17:03:04,614 INFO source.NetcatSource: Created serverSocket:sun.nio.ch.ServerSocketChannelImpl[/127.0.0.1:9999]
```

After this, keep this program running on this terminal and start another terminal and perform the telnet command to write out to our host as follows.

```
mohammed@mohammed-VirtualBox: $ curl telnet://localhost:9999
```

Once this is executed, we can write any messages or texts we want and it will be delivered to the host and be confirmed to us by the “OK” text after each message is passed. Have a look.

```
mohammed@mohammed-VirtualBox:$ curl telnet://localhost:9999
unboxing
OK
big
OK
data
OK
■
```

The texts delivered are displayed on the terminal on which we ran our netcat.conf file.

```
INFO sink.LoggerSink: Event: { headers:{} body: 75 6E 62 6F 78 69 6E 67
INFO sink.LoggerSink: Event: { headers:{} body: 62 69 67
INFO sink.LoggerSink: Event: { headers:{} body: 64 61 74 61
unboxing }
big }
data }
```