Faculty: Ms. Sana Shaikh **Subject:** Big Data Analytics 2020-2021

Name: Alston Fernandes **Roll no:** 19 **Exp no:** 1

Topic:	Installing Hadoop in Pseudo Distributed Mode & getting familiar with Hadoop HDFS commands.
Prerequisite:	Basic knowledge of Hadoop is required.
Mapping With COs:	CSL7012.1
Objective:	• To acquire the knowledge of different components present in Hadoop Ecosystem. • To understand the Hadoop 2.x architecture. • To learn how to set up and configure Hadoop in Pseudo Distributed Mode. • To learn how to work with the Hadoop HDFS file system. • Getting familiar with Hadoop HDFS commands.
Outcome:	• Students will be able to differentiate between the local file system and HDFS. • Students will be able to install Hadoop clusters and configure it in any installation mode. • Students will be able to start and work with the Hadoop HDFS file system using various Hadoop commands.
Instructions:	This experiment is a compulsory experiment. All the students are required to perform this experiment individually.
Deliverables:	**1. Explain all Hadoop installation modes.** **Local (Standalone) Mode:** • This mode is primarily used for development, debugging, and testing purposes. • In this mode, Hadoop runs on a single machine without the need for a distributed cluster. It does not take advantage of Hadoop's distributed computing capabilities. • Only the essential Hadoop components are active, such as HDFS for local file storage and MapReduce for data processing. There is no need for YARN resource management.

- It is easy to set up and suitable for small-scale tasks where distributed computing benefits are unnecessary.

Pseudo-Distributed Mode:

- This mode is often used for learning and development on a single machine that simulates a distributed cluster environment.
- In this mode, all Hadoop components run on a single machine, but they communicate as if they were part of a distributed cluster. It allows developers to test their applications in an environment that resembles a real Hadoop cluster.
- HDFS, YARN, and MapReduce are fully functional.
- Provides a realistic testing environment for Hadoop applications without the complexity of setting up a multi-node cluster.

Cluster (Fully-Distributed) Mode:

- This is the production-ready mode for deploying Hadoop in a distributed cluster, suitable for processing large-scale data.
- In this mode, Hadoop operates in a true distributed cluster environment with multiple nodes. Each node has its HDFS storage and runs various Hadoop services.
- HDFS, YARN, MapReduce, and other Hadoop ecosystem components are distributed across the cluster nodes.
- Offers scalability, fault tolerance, and efficient distributed data processing capabilities for big data workloads.

2. List down components of Hadoop Cluster.

Core Components:

1. **HDFS (Hadoop Distributed File System)**:
2. **YARN (Yet Another Resource Negotiator)**:
3. **MapReduce**:

Optional Components (Hadoop Ecosystem):

1. **Hbase**:
2. **Hive**:
3. **Pig**:
4. **Sqoop**:
5. **Flume**:

6. **Oozie**:
7. **ZooKeeper**:
8. **Mahout**:
9. **Ambari**:

3. Take a snapshot of each step of hadoop installation and for all HDFS commands with input and its output.

1. Update the system

```
dbit@complab3:~$ pwd
/home/dbit
dbit@complab3:~$ sudo apt-get update
[sudo] password for dbit:
Hit:1 http://in.archive.ubuntu.com/ubuntu jammy InRelease
Hit:2 http://security.ubuntu.com/ubuntu jammy-security InRelease
Hit:3 http://in.archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:4 http://in.archive.ubuntu.com/ubuntu jammy-backports InRelease
Reading package lists... Done
dbit@complab3:~$
```

2. Install jdk

```
dbit@complab3: $ 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:
  ca-certificates-java default-jdk-headless default-jre default-jre-headless fonts-dejavu-extra java-common libatk-
  libatk-wrapper-java-jni libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libxau-dev libxcb1-dev libxdmcp-dev
  openjdk-11-jdk openjdk-11-jdk-headless openjdk-11-jre openjdk-11-jre-headless x11proto-dev xorg-sgml-doctools xtr
Suggested packages:
  libice-doc libsm-doc libx11-doc libxcb-doc libxt-doc openjdk-11-demo openjdk-11-source visualvm fonts-ipafont-got
  fonts-wqy-microhei | fonts-wqy-zenhei
The following NEW packages will be installed:
  ca-certificates-java default-jdk default-jdk-headless default-jre default-jre-headless fonts-dejavu-extra java-co
  libatk-wrapper-java-jni libice-dev libpthread-stubs0-dev libsm-dev libx11-dev libxau-dev libxcb1-dev libxdmcp-dev
  openjdk-11-jdk openjdk-11-jdk-headless openjdk-11-jre openjdk-11-jre-headless x11proto-dev xorg-sgml-doctools xtr
0 upgraded, 24 newly installed, 0 to remove and 2 not upgraded.
Need to get 122 MB of archives.
After this operation, 275 MB of additional disk space will be used.
Do you want to continue? [Y/n]
```

```
dbit@complab3: $ java --version
openjdk 11.0.20.1 2023-08-24
OpenJDK Runtime Environment (build 11.0.20.1+1-post-Ubuntu-0ubuntu122.04)
OpenJDK 64-Bit Server VM (build 11.0.20.1+1-post-Ubuntu-0ubuntu122.04, mixed mode, sharing)
dbit@complab3: $
```

3. Install ssh

```
dbit@complab3:~$ sudo apt-get install ssh
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
The following additional packages will be installed:
  ncurses-term openssh-server openssh-sftp-server ssh-import-id
Suggested packages:
  molly-guard monkeysphere ssh-askpass
The following NEW packages will be installed:
  ncurses-term openssh-server openssh-sftp-server ssh ssh-import-id
0 upgraded, 5 newly installed, 0 to remove and 2 not upgraded.
Need to get 755 kB of archives.
After this operation, 6,180 kB of additional disk space will be used.
Do you want to continue? [Y/n]
```

4. Generate ssh key pair

```
dbit@complab3:~$ ssh-keygen -t rsa -P ''
Generating public/private rsa key pair.
Enter file in which to save the key (/home/dbit/.ssh/id_rsa):
Your identification has been saved in /home/dbit/.ssh/id_rsa
Your public key has been saved in /home/dbit/.ssh/id_rsa.pub
The key fingerprint is:
SHA256:LNjr3DIPlykKj12yNBKE7PezBxYvbuVUdjXH7GwHOjw dbit@complab3
The key's randomart image is:
+---[RSA 3072]----+
|              o  |
|..           o.+ |
|...         ...=. |
|..   + . o .E  +.|
| .... = S .  o. .|
|  ...+ * o       |
|  o *o@ +        |
|   B @=B         |
|  . *.++o        |
+----[SHA256]-----+
dbit@complab3:~$
```

5. Connect to localhost via ssh

```
dbit@complab3:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ED25519 key fingerprint is SHA256:fEyXPWreVyvn45GMOvwmfpP6IlOJjVwbU1z/JHcB9n8.
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.
dbit@localhost's password:
Welcome to Ubuntu 22.04.3 LTS (GNU/Linux 6.2.0-31-generic x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

Expanded Security Maintenance for Applications is not enabled.

0 updates can be applied immediately.

Enable ESM Apps to receive additional future security updates.
See https://ubuntu.com/esm or run: sudo pro status


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

Ubuntu comes with ABSOLUTELY NO WARRANTY, to the extent permitted by
applicable law.

dbit@complab3:~$
```

6. Move hadoop package to usr folder

```
dbit@complab3:~$ sudo mv ~/Downloads/hadoop-2.7.7 /usr/local/hadoop
[sudo] password for dbit:
dbit@complab3:~$
```

7. Add environment variables to bashrc file and update it

```
 GNU nano 6.2                                    .bashrc *
#   sleep 10; alert
alias alert='notify-send --urgency=low -i "$([ $? = 0 ] && echo terminal || echo error)" "$(history|tail -n1|sed -e

# 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


#Hadoop variables
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
export HADOOP_INSTALL=/usr/local/hadoop
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
#end of Hadoop variable declaration
```

```
dbit@complab3:~$ nano .bashrc
dbit@complab3:~$ source .bashrc
dbit@complab3:~$
```

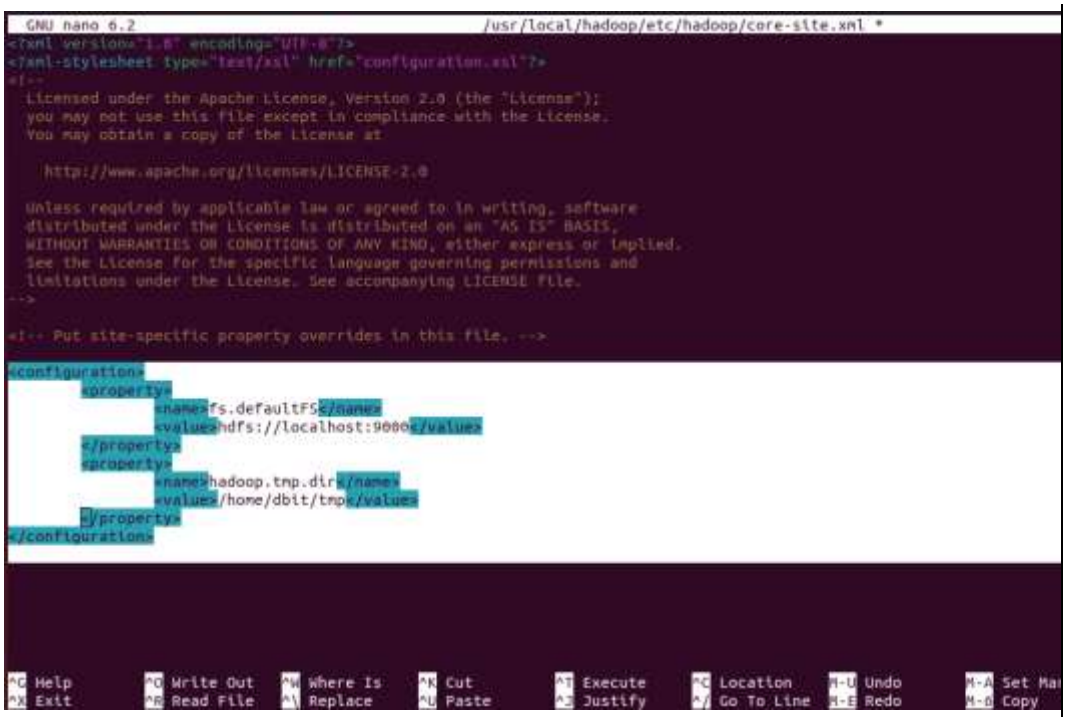
8. Check hadoop version

```
dbit@complab3:~$ hadoop version
Hadoop 2.7.7
Subversion Unknown -r c1aad84bd27cd79c3d1a7dd58202a8c3ee1ed3ac
Compiled by stevel on 2018-07-18T22:47Z
Compiled with protoc 2.5.0
From source with checksum 792e15d20b12c74bd6f19a1fb886490
This command was run using /usr/local/hadoop/share/hadoop/common/hadoop-common-2.7.7.jar
dbit@complab3:~$
```

9. Update hadoop environment settings

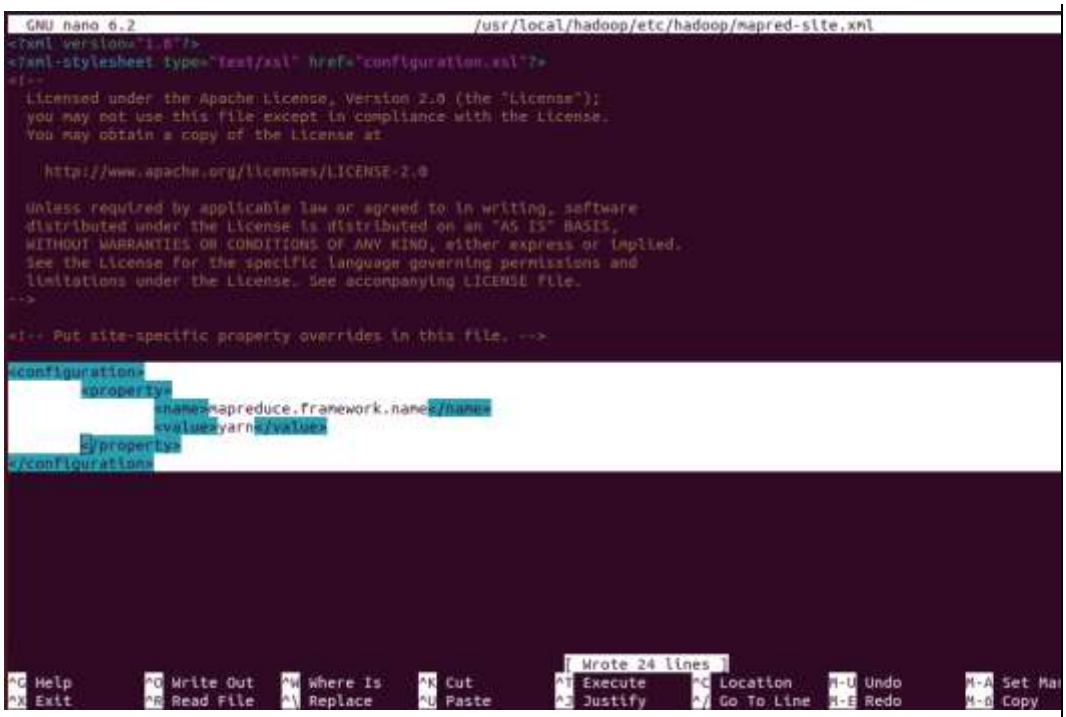
```
  GNU nano 6.2                          /usr/local/hadoop/etc/hadoop/hadoop-env.sh *
# Licensed to the Apache Software Foundation (ASF) under one
# or more contributor license agreements.  See the NOTICE file
# distributed with this work for additional information
# regarding copyright ownership.  The ASF licenses this file
# to you 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.

# 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=usr/lib/jvm/java-1.8.0-openjdk-amd64

# 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"}

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

10. Update hadoop core file

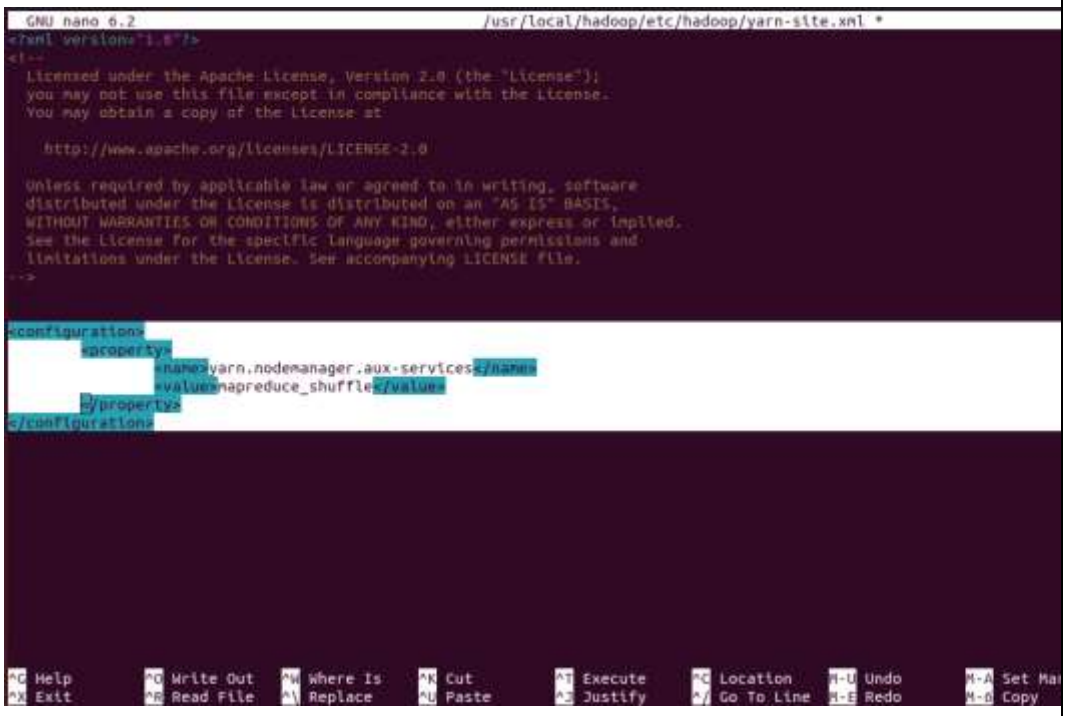


11. Update hadoop hdfs file

12. Update hadoop map reduce file



13. Update hadoop yarn file

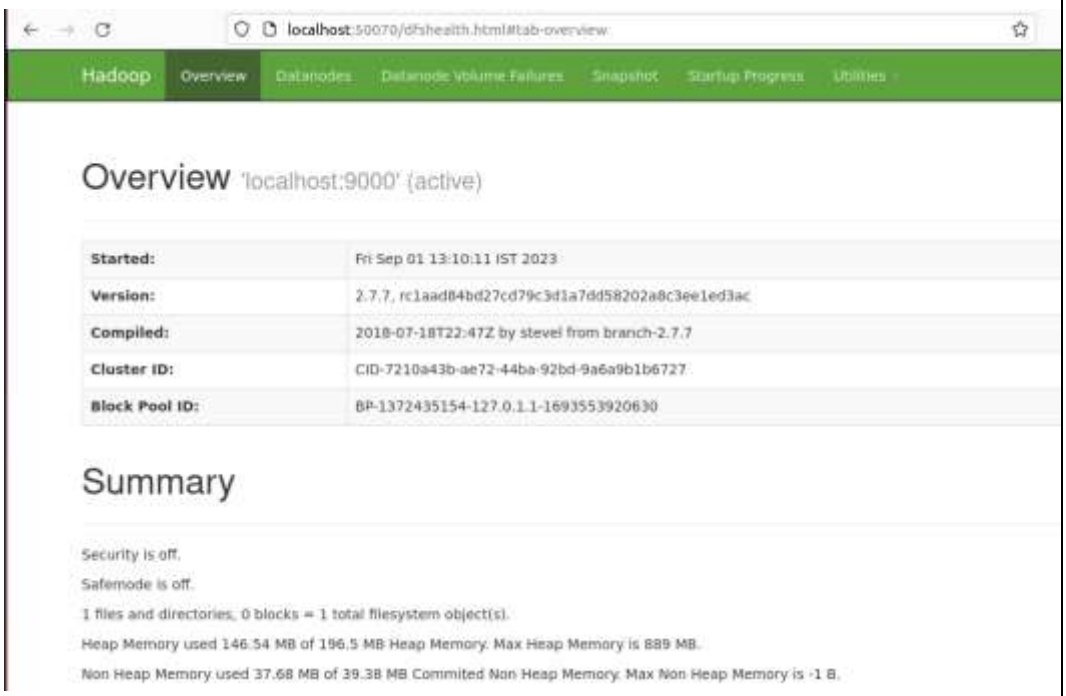


14. Starting hadoop

```
dbit@complab3:~$ start-dfs.sh
Starting namenodes on [localhost]
dbit@localhost's password:
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-dbit-namenode-complab3.out
dbit@localhost's password:
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-dbit-datanode-complab3.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:fEyXPWreVyvn45GMDvwmfpP6IlOJjVwbU1z/JHcB9n8.
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.
dbit@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-dbit-secondarynamenode-complab3.out
dbit@complab3:~$ start-yarn.sh
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-dbit-resourcemanager-complab3.out
dbit@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-dbit-nodemanager-complab3.out
dbit@complab3:~$
```

```
dbit@complab3:~$ jps
11062 SecondaryNameNode
10742 NameNode
10889 DataNode
11660 Jps
11197 ResourceManager
11486 NodeManager
dbit@complab3:~$
```

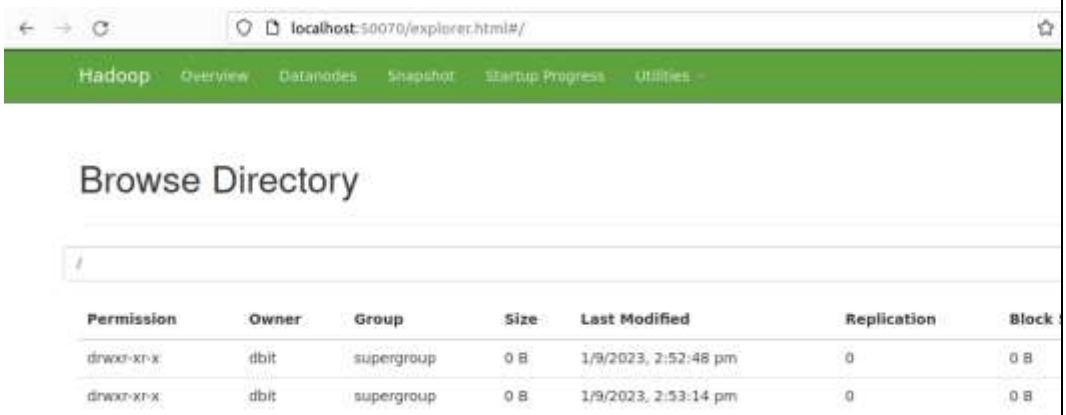
15. Hadoop dashboard



16. Basic hadoop commands

mkdir

```
dbit@complab3:~$ hadoop fs -mkdir /bda
dbit@complab3:~$ hadoop fs -mkdir /bda
mkdir: `/bda': File exists
dbit@complab3:~$ hadoop fs -mkdir /test
dbit@complab3:~$
```



copy and move files

```
dbit@complab3:~$ hadoop fs -put Downloads/ExperimentNo1.pdf /bda/
dbit@complab3:~$ hadoop fs -moveFromLocal Downloads/ExperimentNo1.pdf /test/
dbit@complab3:~$
```

ls command

```
dbit@complab3:~$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /bda
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /test
dbit@complab3:~$ hadoop fs -ls /bda
Found 1 items
-rw-r--r--   1 dbit supergroup     108614 2023-09-01 18:28 /bda/ExperimentNo1.pdf
dbit@complab3:~$ hadoop fs -ls /test
Found 1 items
-rw-r--r--   1 dbit supergroup     108614 2023-09-01 18:28 /test/ExperimentNo1.pdf
dbit@complab3:~$
```

get command

```
dbit@complab3:~$ hadoop fs -get /test/ExperimentNo1.pdf ~/Downloads/
dbit@complab3:~$ ls ~/Downloads/
ExperimentNo1.pdf  hadoop-2.7.7.tar.gz
dbit@complab3:~$
```

touch command

```
dbit@complab3:~$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /bda
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /test
dbit@complab3:~$ hadoop fs -touch hi.txt
-touch: Unknown command
dbit@complab3:~$ hadoop fs -touchz hi.txt
touchz: `hi.txt': No such file or directory
dbit@complab3:~$ hadoop fs -touchz /hi.txt
dbit@complab3:~$ hadoop fs -ls /
Found 3 items
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /bda
-rw-r--r--   1 dbit supergroup          0 2023-09-03 19:42 /hi.txt
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /test
dbit@complab3:~$
```

cat command

```
dbit@complab3:~$ hadoop fs -cat /test/hi1.txt
Hello
dbit@complab3:~$ hadoop fs -cat /test/hi2.txt
, World!
dbit@complab3:~$
```

cp and mv command

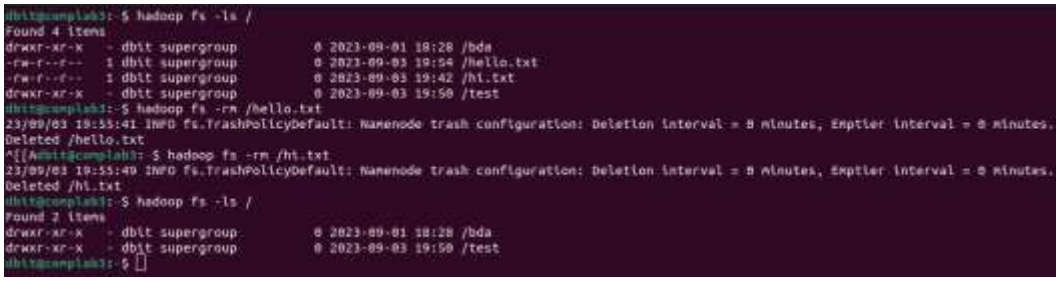
```
dbit@complab3:~$ hadoop fs -ls /
Found 3 items
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /bda
-rw-r--r--   1 dbit supergroup         33 2023-09-03 19:56 /hello.txt
drwxr-xr-x   - dbit supergroup          0 2023-09-03 19:50 /test
dbit@complab3:~$ hadoop fs -ls /test
Found 5 items
-rw-r--r--   1 dbit supergroup     108614 2023-09-01 18:28 /test/ExperimentNo1.pdf
-rw-r--r--   1 dbit supergroup          6 2023-09-03 19:50 /test/hi1.txt
-rw-r--r--   1 dbit supergroup          9 2023-09-03 19:50 /test/hi2.txt
-rw-r--r--   1 dbit supergroup         10 2023-09-03 19:50 /test/hi3.txt
-rw-r--r--   1 dbit supergroup          8 2023-09-03 19:50 /test/hi4.txt
dbit@complab3:~$ hadoop fs -ls /bda
Found 1 items
-rw-r--r--   1 dbit supergroup     108614 2023-09-01 18:28 /bda/ExperimentNo1.pdf
dbit@complab3:~$ hadoop fs -cp /hello.txt /bda
dbit@complab3:~$ hadoop fs -ls /
Found 3 items
drwxr-xr-x   - dbit supergroup          0 2023-09-03 19:58 /bda
-rw-r--r--   1 dbit supergroup         33 2023-09-03 19:56 /hello.txt
drwxr-xr-x   - dbit supergroup          0 2023-09-03 19:50 /test
dbit@complab3:~$ hadoop fs -ls /bda
Found 2 items
-rw-r--r--   1 dbit supergroup     108614 2023-09-01 18:28 /bda/ExperimentNo1.pdf
-rw-r--r--   1 dbit supergroup         33 2023-09-03 19:58 /bda/hello.txt
dbit@complab3:~$ hadoop fs -mv /hello.txt /test
dbit@complab3:~$ hadoop fs -ls /test
Found 6 items
-rw-r--r--   1 dbit supergroup     108614 2023-09-01 18:28 /test/ExperimentNo1.pdf
-rw-r--r--   1 dbit supergroup         33 2023-09-03 19:56 /test/hello.txt
-rw-r--r--   1 dbit supergroup          6 2023-09-03 19:50 /test/hi1.txt
-rw-r--r--   1 dbit supergroup          9 2023-09-03 19:50 /test/hi2.txt
-rw-r--r--   1 dbit supergroup         10 2023-09-03 19:50 /test/hi3.txt
-rw-r--r--   1 dbit supergroup          8 2023-09-03 19:50 /test/hi4.txt
```

```
dbit@complab3:~$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - dbit supergroup          0 2023-09-03 19:58 /bda
drwxr-xr-x   - dbit supergroup          0 2023-09-03 19:59 /test
dbit@complab3:~$
```

appendToFile command

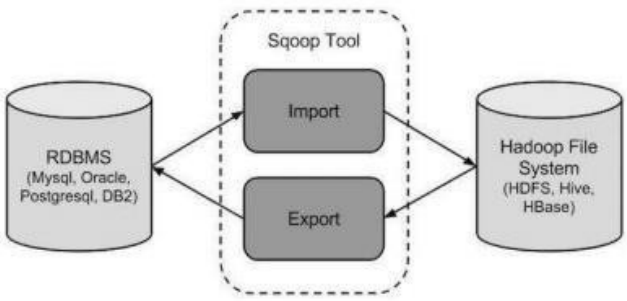
```
dbit@complab3:~$ hadoop fs -ls /
Found 2 items
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /bda
drwxr-xr-x   - dbit supergroup          0 2023-09-03 19:50 /test
dbit@complab3:~$ hadoop fs -appendToFile hi1.txt hi2.txt hi3.txt hi4.txt /hello.txt
dbit@complab3:~$ hadoop fs -ls /
Found 3 items
drwxr-xr-x   - dbit supergroup          0 2023-09-01 18:28 /bda
-rw-r--r--   1 dbit supergroup         33 2023-09-03 19:56 /hello.txt
drwxr-xr-x   - dbit supergroup          0 2023-09-03 19:50 /test
dbit@complab3:~$ hadoop fs -cat /hello.txt
Hello
, World!
 This is
Alston.
dbit@complab3:~$
```

rm command

	
Conclusion:	I was able to set up and configure Hadoop cluster in Pseudo Distributed Mode, which helps to simulate a multi node installation on a single node and also will able to work with Hadoop HDFS file system using various commands.
References:	https://data-flair.training/blogs/install-hadoop-on-ubuntu/

Experiment No: 2

Name: Alston Fernandes **Roll No: 19** **Batch: A**

Topic:	Use of Sqoop tool to transfer data between Hadoop and relational databaseservers. a. Sqoop and MySQL - Installation. To execute basic commands of Hadoop eco system component Sqoop.
Prerequisite:	o Familiarity with command-line interfaces such as bash o Basic knowledge of Relational database management systems. MySQL Basic familiarity with the purpose and operation of Hadoop
	o
Mapping With COs:	CSL704.3
Objective:	Ingest data using Sqoop.
Outcome:	Students will be able to use the Sqoop tool - for transferring data between Hadoop &relational databases
Instructions:	This experiment is a compulsory experiment. All the students are required to perform this experiment individually.
Deliverables:	**SQOOP INSTALLATION** 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. 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, isstored 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 tointeract 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 databaseserver 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 asMySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases. It is provided by the Apache Software Foundation. The following image describes the workflow of Sqoop. 

Sqoop Import
The import tool imports individual tables from RDBMS to HDFS. Each row ina table is treated as a record in HDFS. All records are stored as text data in text files or as binary data in Avro and Sequence files.

Sqoop Export
The export tool exports a set of files from HDFS back to an RDBMS. The filesgiven 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 userspecified delimiter.

STEPS to install Sqoop : 1.

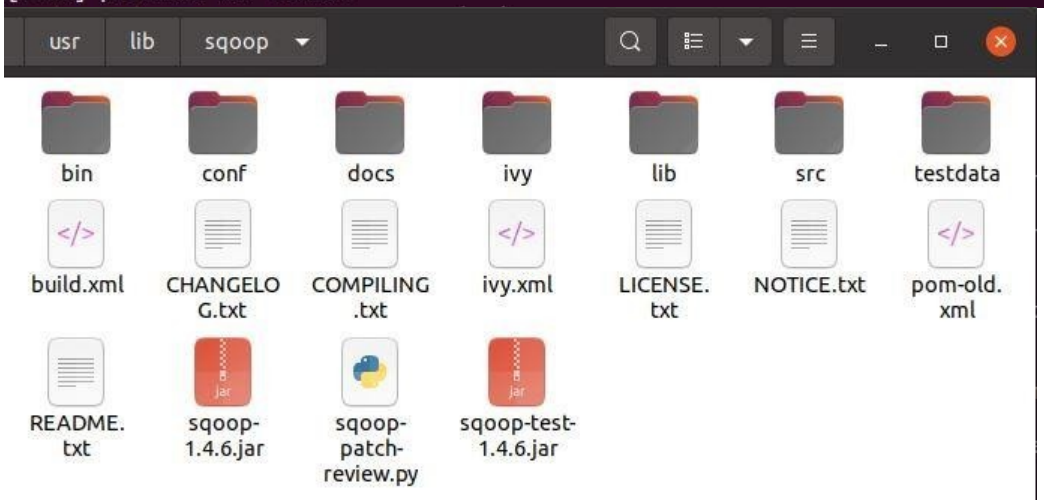
1.Extract the Sqoop Package from the tar file pasted on the Desktop. The extractedpackage can be seen, listed in the list of files and folders of the Desktop using the ls command.

ls Desktop

```
slade@slade-VirtualBox:~$ jps
3937 Jps
3844 NodeManager
3284 NameNode
3575 SecondaryNameNode
3417 DataNode
3708 ResourceManager
slade@slade-VirtualBox:~$ ls Desktop
sqoop-1.4.6.bin__hadoop-2.0.4-alpha
```

2. Move this extracted folder (sqoop1.4.6.bin_hadoop2.0.4alpha) fromDesktop to the directory /usr/lib/sqoop using the sudo mv command.

```
slade@slade-VirtualBox:~$ sudo mv /home/slade/Desktop/sqoop-1.4.6.bin__hadoop-2
alpha /usr/lib/sqoop
[sudo] password for slade:
```



3. Sqoop environment can be set up only by appending the following linesby executing nano ~/.bashrc command.

```
slade@slade-VirtualBox:~$ nano ~/.bashrc
```

Append the following lines in this file.

export SQOOP_HOME=/usr/lib/sqoop
export PATH=$PATH:
$SQOOP_HOME/bin
Ctrl+X ....Y. .. Enter

```
#Hadoop variables
export JAVA_HOME=/usr/lib/jvm/java-1.8.0-openjdk-amd64
export HADOOP_INSTALL=/usr/local/hadoop
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 SQOOP_HOME=/usr/lib/sqoop
export PATH=$PATH:$SQOOP_HOME/bin
#End of Hadoop variable declaration
```

4. Now save this bashrc file permanently by the command **source ~/.bashrc**

```
slade@slade-VirtualBox:~$ source ~/.bashrc
```

5. To configure Sqoop with Hadoop we need to edit a file sqoopenv.sh whichis present in the directory path $SQOOP_HOME/conf.

```
slade@slade-VirtualBox:~$ cd $SQOOP_HOME/conf
slade@slade-VirtualBox:/usr/lib/sqoop/conf$ ls
oraoop-site-template.xml  sqoop-env-template.sh    sqoop-site.xml
sqoop-env-template.cmd    sqoop-site-template.xml
```

Now move the contents of the template file sqoopenv-template.sh to sqoopenv.sh using the mv command. mv sqoopenvtemplate.sh sqoop-env.sh

```
slade@slade-VirtualBox:/usr/lib/sqoop/conf$ mv sqoop-env-template.sh sqoop-env.sh
slade@slade-VirtualBox:/usr/lib/sqoop/conf$ ls
oraoop-site-template.xml  sqoop-env-template.cmd   sqoop-site.xml
sqoop-env.sh              sqoop-site-template.xml
```

To add contents in the Sqoop-env.sh file use the command nano sqoop-env.sh

```
slade@slade-VirtualBox:/usr/lib/sqoop/conf$ nano sqoop-env.sh
```

slade@slade-VirtualBox: /usr/lib/sqoop/conf

```
 GNU nano 4.8                        sqoop-env.sh                        Modified
# 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.

# included in all the hadoop scripts with source command
# should not be executable directly
# also should not be passed any arguments, since we need original $*

# Set Hadoop-specific environment variables here.

#Set path to where bin/hadoop is available
export HADOOP_COMMON_HOME=/usr/local/hadoop

#Set path to where hadoop-*-core.jar is available
export HADOOP_MAPRED_HOME=/usr/local/hadoop

#set the path to where bin/hbase is available
#export HBASE_HOME=

#Set the path to where bin/hive is available
#export HIVE_HOME=

#Set the path for where zookeper config dir is
#export ZOOCFGDIR=
```

Crtl+X...Y ...Enter

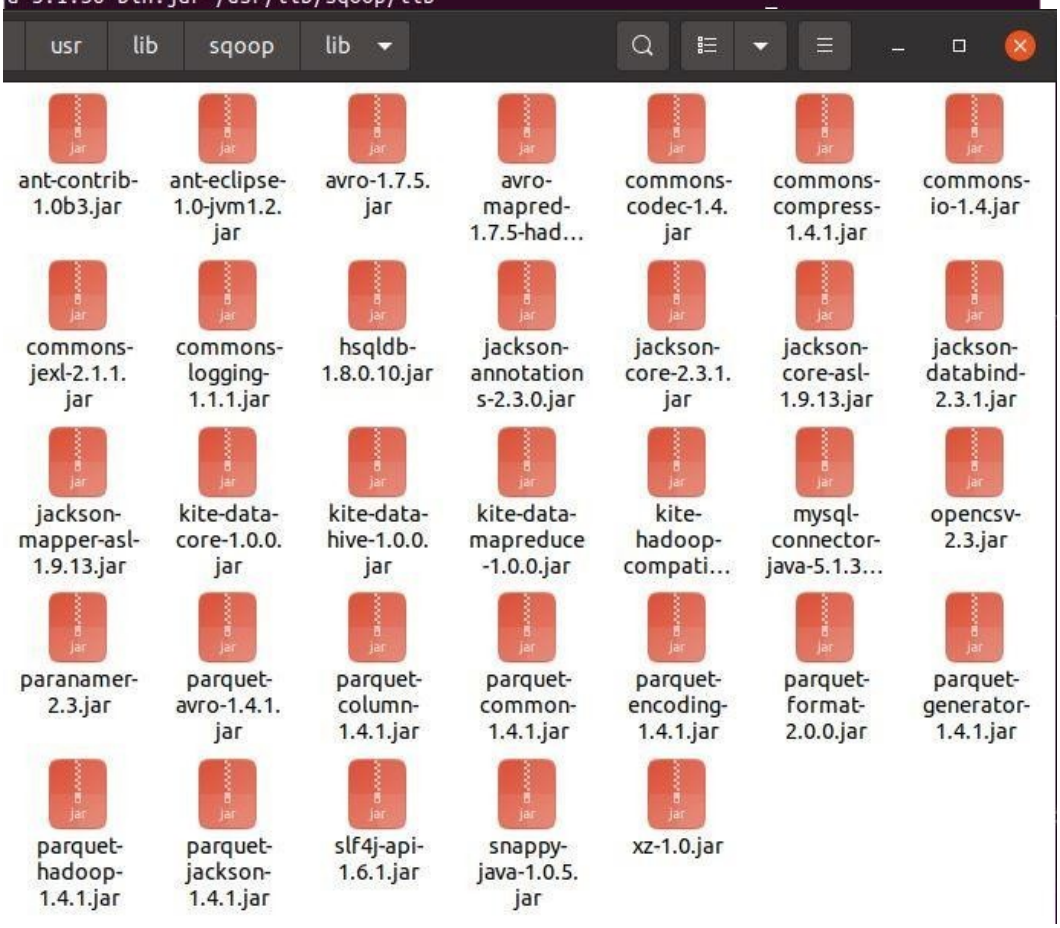
6. Now copy, add or download the mysqlconnectorjava5.1.36.tar.gz fileonto the Desktop. Extract this file in the same file location.

```
slade@slade-VirtualBox:/usr/lib/sqoop/conf$ cd
slade@slade-VirtualBox:~$ cd Desktop
slade@slade-VirtualBox:~/Desktop$ ls
mysql-connector-java-5.1.36
```

7. Move this extracted file to the location /usr/lib/sqoop/lib using the mv command.

cd Desktop
ls
cd mysql•connector•java•5.1.36
ls
mv mysql•connector•java•5.1.36•bin.jar /usr/lib/sqoop/lib
ls /usr/lib/sqoop/lib

```
slade@slade-VirtualBox:~/Desktop$ cd mysql-connector-java-5.1.36
slade@slade-VirtualBox:~/Desktop/mysql-connector-java-5.1.36$ ls
build.xml  COPYING  mysql-connector-java-5.1.36-bin.jar  README.txt
CHANGES    docs     README                                src
slade@slade-VirtualBox:~/Desktop/mysql-connector-java-5.1.36$ mv mysql-connector-jav
a-5.1.36-bin.jar /usr/lib/sqoop/lib
```



8. To check if Sqoop has been installed correctly we move to the directory $SQOOP_HOME/bin and use the command sqoop version to check for sqoop installation success.

cd /usr/lib/sqoop/bin
cd $SQOOP_HOME/bin
sqoop version

```
slade@slade-VirtualBox:~/Desktop/mysql-connector-java-5.1.36$ cd
slade@slade-VirtualBox:~$ cd /usr/lib/sqoop/bin
slade@slade-VirtualBox:/usr/lib/sqoop/bin$ cd $SQOOP_HOME/bin
slade@slade-VirtualBox:/usr/lib/sqoop/bin$ sqoop version
Warning: /usr/lib/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/lib/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
21/08/17 19:56:47 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
Sqoop 1.4.6
git commit id c0c5a81723759fa575844a0a1eae8f510fa32c25
Compiled by root on Mon Apr 27 14:38:36 CST 2015
```

STEPS for MYSQL INSTALLATION

1. After Sqoop installation MySQL has to be installed as well. Firstly, installall the required libraries using the command sudo apt•get install mysql•server.

```
slade@slade-VirtualBox:~$ sudo apt-get install mysql-server
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following packages were automatically installed and are no longer required:
  libllvm7 libllvm9 linux-hwe-5.4-headers-5.4.0-53 linux-hwe-5.4-headers-5.4.0-65 linux-hwe-5.4-headers-5.4.0-71
  linux-hwe-5.4-headers-5.4.0-74 python3-click python3-colorama
Use 'sudo apt autoremove' to remove them.
The following additional packages will be installed:
  libaio1 libevent-core-2.1-6 libhtml-template-perl mysql-client-5.7 mysql-client-core-5.7 mysql-server-5.7 mysql-server-core-5.7
Suggested packages:
  libipc-sharedcache-perl mailx tinyca
The following NEW packages will be installed:
  libaio1 libevent-core-2.1-6 libhtml-template-perl mysql-client-5.7 mysql-client-core-5.7 mysql-server mysql-server-5.7
  mysql-server-core-5.7
0 upgraded, 8 newly installed, 0 to remove and 101 not upgraded.
Need to get 19.1 MB of archives.
After this operation, 154 MB of additional disk space will be used.
Do you want to continue? [Y/n]
```

```
Selecting previously unselected package mysql-client-core-5.7.
Preparing to unpack .../1-mysql-client-core-5.7_5.7.35-0ubuntu0.18.04.1_amd64.deb ...
Unpacking mysql-client-core-5.7 (5.7.35-0ubuntu0.18.04.1) ...
Selecting previously unselected package mysql-client-5.7.
Preparing to unpack .../2-mysql-client-5.7_5.7.35-0ubuntu0.18.04.1_amd64.deb ...
Unpacking mysql-client-5.7 (5.7.35-0ubuntu0.18.04.1) ...
Selecting previously unselected package mysql-server-core-5.7.
Preparing to unpack .../3-mysql-server-core-5.7_5.7.35-0ubuntu0.18.04.1_amd64.deb ...
Unpacking mysql-server-core-5.7 (5.7.35-0ubuntu0.18.04.1) ...
Selecting previously unselected package libevent-core-2.1-6:amd64.
Preparing to unpack .../4-libevent-core-2.1-6_2.1.8-stable-4build1_amd64.deb ...
Unpacking libevent-core-2.1-6:amd64 (2.1.8-stable-4build1) ...
Selecting previously unselected package mysql-server-5.7.
Preparing to unpack .../5-mysql-server-5.7_5.7.35-0ubuntu0.18.04.1_amd64.deb ...
Unpacking mysql-server-5.7 (5.7.35-0ubuntu0.18.04.1) ...
Selecting previously unselected package libhtml-template-perl.
Preparing to unpack .../6-libhtml-template-perl_2.97-1_all.deb ...
Unpacking libhtml-template-perl (2.97-1) ...
Selecting previously unselected package mysql-server.
Preparing to unpack .../7-mysql-server_5.7.35-0ubuntu0.18.04.1_all.deb ...
Unpacking mysql-server (5.7.35-0ubuntu0.18.04.1) ...
Setting up libevent-core-2.1-6:amd64 (2.1.8-stable-4build1) ...
Setting up libhtml-template-perl (2.97-1) ...
Setting up libaio1:amd64 (0.3.110-5ubuntu0.1) ...
Setting up mysql-client-core-5.7 (5.7.35-0ubuntu0.18.04.1) ...
Setting up mysql-server-core-5.7 (5.7.35-0ubuntu0.18.04.1) ...
Setting up mysql-client-5.7 (5.7.35-0ubuntu0.18.04.1) ...
Setting up mysql-server-5.7 (5.7.35-0ubuntu0.18.04.1) ...
update-alternatives: using /etc/mysql/mysql.cnf to provide /etc/mysql/my.cnf (my.cnf) in auto mode
Renaming removed key_buffer and myisam-recover options (if present)
Created symlink /etc/systemd/system/multi-user.target.wants/mysql.service → /lib/systemd/system/mysql.service.
Setting up mysql-server (5.7.35-0ubuntu0.18.04.1) ...
Processing triggers for libc-bin (2.27-3ubuntu1.4) ...
Processing triggers for systemd (237-3ubuntu10.50) ...
Processing triggers for man-db (2.8.3-2ubuntu0.1) ...
Processing triggers for ureadahead (0.100.0-21) ...
ureadahead will be reprofiled on next reboot
```

3. To login to the MySQL user, use the following command: mysql u root p

```
slade@slade-VirtualBox:~$ sudo mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 14
Server version: 5.7.35-0ubuntu0.18.04.1 (Ubuntu)

Copyright (c) 2000, 2021, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

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

mysql>
```

It will be asked to enter the password for the corresponding user. Enter the password. Now the MySQL script will run and the user will be logged in. This verifies the successful completion of the MySQL installation onto the system.

IMPORT/EXPORT

1. We check if all the services are running using the jps command.

```
slade@slade-VirtualBox:~$ jps
3937 Jps
3844 NodeManager
3284 NameNode
3575 SecondaryNameNode
3417 DataNode
3708 ResourceManager
```

2. Then we start the mysql shell as:

```
slade@slade-VirtualBox:~$ sudo mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 14
Server version: 5.7.35-0ubuntu0.18.04.1 (Ubuntu)

Copyright (c) 2000, 2021, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

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

mysql>
```

3. We see the list of databases.

```
mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| mysql              |
| performance_schema |
| sys                |
+--------------------+
4 rows in set (0.00 sec)
```

4. We create a new database or use an existing database according to the need.

```
mysql> create database emp;
Query OK, 1 row affected (0.00 sec)

mysql> show databases;
+--------------------+
| Database           |
+--------------------+
| information_schema |
| emp                |
| mysql              |
| performance_schema |
| sys                |
+--------------------+
5 rows in set (0.01 sec)
```

5. Now we create a table in mysql which we will import into HDFS. create table Faculty (id int primary key, name varchar(10), city varchar(10), salary bigint);

```
mysql> create table Faculty (id int primary key, name varchar(10), city
    -> varchar(10), salary bigint);
Query OK, 0 rows affected (0.30 sec)

mysql> show tables;
+---------------+
| Tables_in_emp |
+---------------+
| Faculty       |
+---------------+
1 row in set (0.00 sec)
```

mysql> Insert into Faculty values(1, 'Sana', 'Mumbai', 95000);
mysql> Insert into Faculty values(2, 'Riya', 'Pune', 85000);
mysql> Insert into Faculty values(3, 'Karan', 'Jaipur', 55000);
mysql> Insert into Faculty values(4, 'Rahul', 'Delhi', 78000);
mysql> Insert into Faculty values(5, 'Bush', 'Mumbai',
75000); mysql> Insert into Faculty values(6, 'Ram', 'Delhi',
66000); mysql> Insert into Faculty values(7, 'Slade', 'Pune',
71000);

```
mysql> Insert into Faculty values(1, 'Sana', 'Mumbai', 95000);
Query OK, 1 row affected (0.27 sec)

mysql> Insert into Faculty values(2, 'Riya', 'Pune', 85000);
Query OK, 1 row affected (0.10 sec)

mysql> Insert into Faculty values(3, 'Karan', 'Jaipur', 55000);
Query OK, 1 row affected (0.12 sec)

mysql> Insert into Faculty values(4, 'Rahul', 'Delhi', 78000);
Query OK, 1 row affected (0.08 sec)

mysql> Insert into Faculty values(5, 'Bush', 'Mumbai', 75000);
Query OK, 1 row affected (0.14 sec)

mysql> Insert into Faculty values(6, 'Ram', 'Delhi', 66000);
Query OK, 1 row affected (0.14 sec)

mysql> Insert into Faculty values(7, 'Slade', 'Pune', 71000);
Query OK, 1 row affected (0.27 sec)
```

6. We can output the entries of the table as follows:

```
mysql> select * from Faculty;
+----+-------+--------+--------+
| id | name  | city   | salary |
+----+-------+--------+--------+
|  1 | Sana  | Mumbai |  95000 |
|  2 | Riya  | Pune   |  85000 |
|  3 | Karan | Jaipur |  55000 |
|  4 | Rahul | Delhi  |  78000 |
|  5 | Bush  | Mumbai |  75000 |
|  6 | Ram   | Delhi  |  66000 |
|  7 | Slade | Pune   |  71000 |
+----+-------+--------+--------+
7 rows in set (0.05 sec)
```

8. Now, we grant privileges to the user so that we can perform import function. grantall privileges on *.* to ‘root’@’localhost’;

```
mysql> grant all privileges on *.* to 'root'@'localhost';
Query OK, 0 rows affected (0.00 sec)
```

9. After that we quit the mysql shell.

```
mysql> quit
Bye
```

TRANSFERRING AN ENTIRE TABLE INTO HADOOP:

10. The command is as follows: sqoop import --connect jdbc:mysql://127.0.0.1:3306/emp --username root --password mySQL12345 --table Faculty -m 1

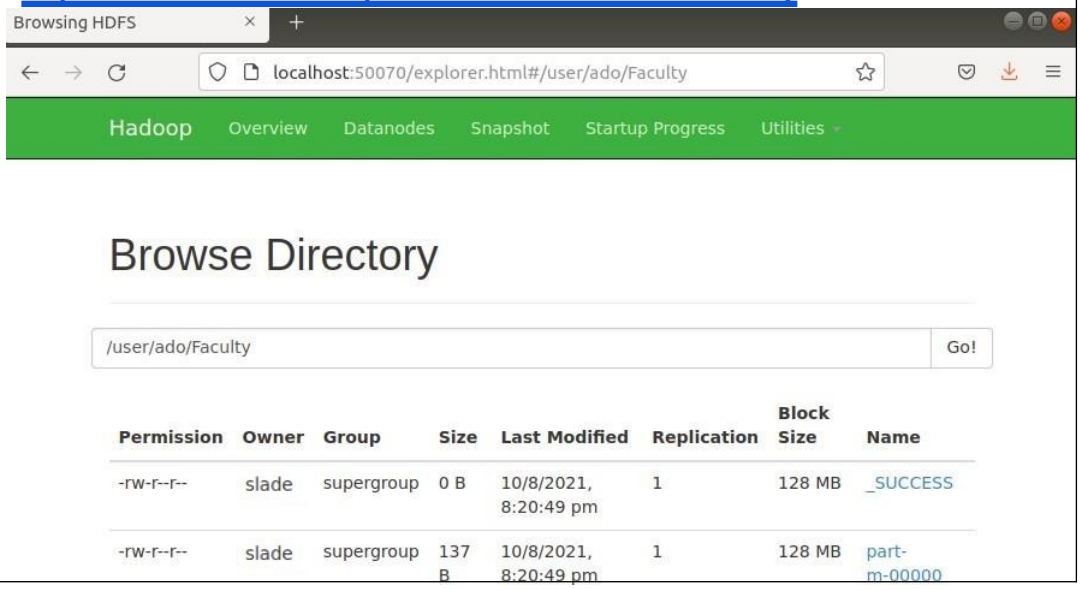
hadoop fs -ls Faculty

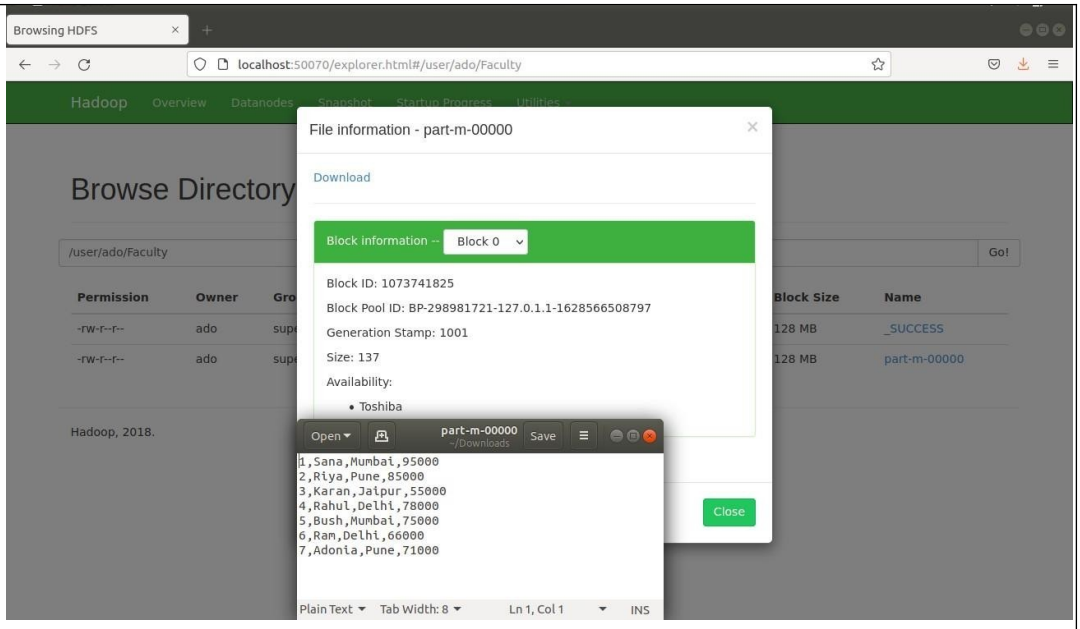
```
slade@slade-VirtualBox:~$ hadoop fs -ls Faculty
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil
(file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.7.7.jar) to method sun.security.krb5
.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentic
ation.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operati
ons
WARNING: All illegal access operations will be denied in a future release
Found 2 items
-rw-r--r--   1 slade supergroup          0 2021-08-10 20:20 Faculty/_SUCCESS
-rw-r--r--   1 slade supergroup        137 2021-08-10 20:20 Faculty/part-m-00000
```

We can check the output using the cat command.

```
slade@slade-VirtualBox:~$ hadoop fs -cat Faculty/part-m-00000
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil
(file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.7.7.jar) to method sun.security.krb5
.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop.security.authentic
ation.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operati
ons
WARNING: All illegal access operations will be denied in a future release
1,Sana,Mumbai,95000
2,Riya,Pune,85000
3,Karan,Jaipur,55000
4,Rahul,Delhi,78000
5,Bush,Mumbai,75000
6,Ram,Delhi,66000
7,Slade,Pune,71000
```

http://localhost:50070/explorer.html#/user/slade/Faculty





TRANSFERRING AN ENTIRE TABLE INTO HADOOP:

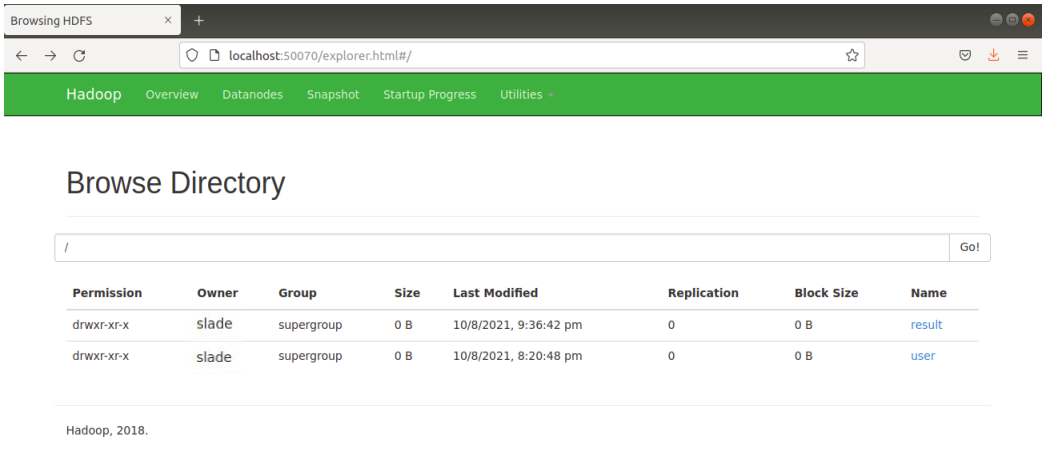
. The command is as follows:

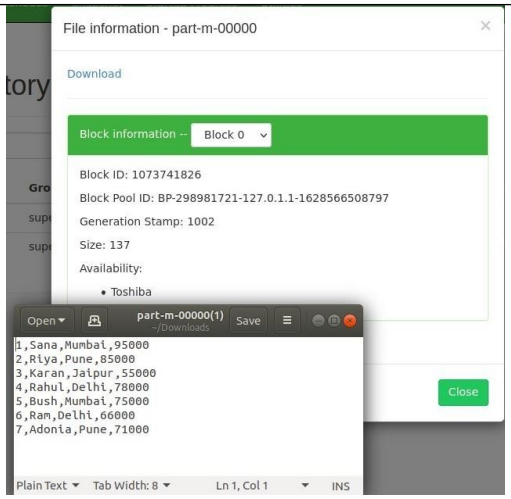
sqoop import –-connect jdbc:mysql://127.0.0.1:3306/Emp –-username root –-password root –-table Faculty -m 1

If –m 1 is not used then the output is not saved in a single partition. It makes more than 1 partitions in the hdfs.

```
ado@Toshiba:~$ sqoop import --connect jdbc:mysql://127.0.0.1:3306/emp ado@Toshiba:~$ sqoop import --connect jdbc:mysql://127.0.0.1:3306/emp --
username root --password Adonia@12 --table Faculty -m 1 --target-dir /result
Warning: /usr/lib/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/lib/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
21/08/10 21:36:38 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
21/08/10 21:36:38 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
21/08/10 21:36:38 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
21/08/10 21:36:38 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 reg
istered via the SPI and manual loading of the driver class is generally unnecessary.
21/08/10 21:36:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Faculty` AS t LIMIT 1
21/08/10 21:36:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Faculty` AS t LIMIT 1
21/08/10 21:36:38 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-ado/compile/77876d659b9ba1eecc1dd4338e1eec3a/Faculty.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
21/08/10 21:36:39 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ado/compile/77876d659b9ba1eecc1dd4338e1eec3a/Faculty.jar
21/08/10 21:36:39 WARN manager.MySQLManager: It looks like you are importing from mysql.
21/08/10 21:36:39 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
21/08/10 21:36:39 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
21/08/10 21:36:39 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
21/08/10 21:36:39 INFO mapreduce.ImportJobBase: Beginning import of Faculty
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/
lib/hadoop-auth-2.7.7.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
21/08/10 21:36:40 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
21/08/10 21:36:40 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
21/08/10 21:36:40 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
21/08/10 21:36:40 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
```

```
                Total committed heap usage (bytes)=132120576
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=137
21/08/10 21:36:42 INFO mapred.LocalJobRunner: Finishing task: attempt_local1884076220_0001_m_000000_0
21/08/10 21:36:42 INFO mapred.LocalJobRunner: map task executor complete.
21/08/10 21:36:42 INFO mapreduce.Job: Job job_local1884076220_0001 running in uber mode : false
21/08/10 21:36:42 INFO mapreduce.Job:  map 100% reduce 0%
21/08/10 21:36:42 INFO mapreduce.Job: Job job_local1884076220_0001 completed successfully
21/08/10 21:36:42 INFO mapreduce.Job: Counters: 20
        File System Counters
                FILE: Number of bytes read=19489276
                FILE: Number of bytes written=19950174
                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=137
                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=7
                Map output records=7
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=3
                Total committed heap usage (bytes)=132120576
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=137
21/08/10 21:36:42 INFO mapreduce.ImportJobBase: Transferred 137 bytes in 2.0881 seconds (65.6094 bytes/sec)
21/08/10 21:36:42 INFO mapreduce.ImportJobBase: Retrieved 7 records.
```



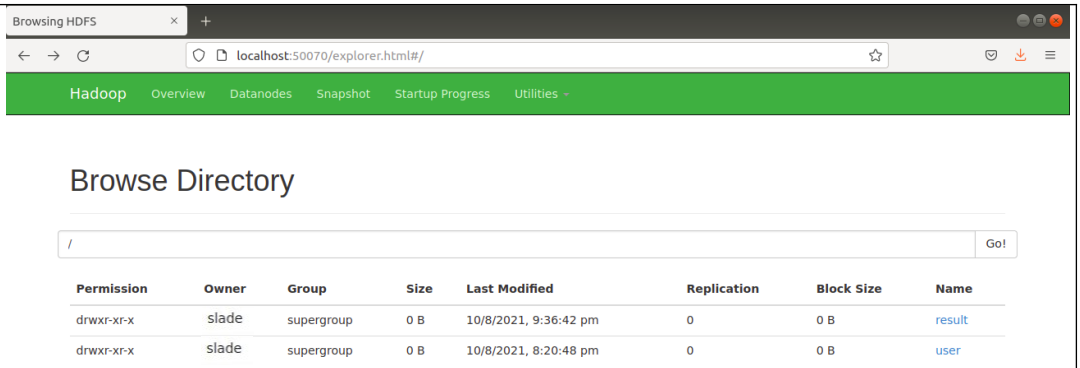


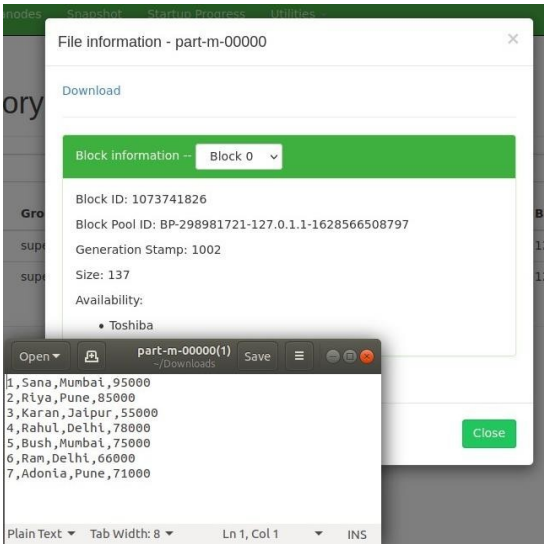
Exercises for Students:

SPECIFYING A TARGET DIRECTORY:

1. Specify the target directory in HDFS into which we want the output tobe saved.

```
Warning: /usr/lib/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/lib/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
21/08/10 21:36:38 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
21/08/10 21:36:38 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
21/08/10 21:36:38 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
21/08/10 21:36:38 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 re
istered via the SPI and manual loading of the driver class is generally unnecessary.
21/08/10 21:36:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Faculty` AS t LIMIT 1
21/08/10 21:36:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Faculty` AS t LIMIT 1
21/08/10 21:36:38 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-ado/compile/77876d659b9ba1eecc1dd4338e1eec3a/Faculty.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
21/08/10 21:36:39 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ado/compile/77876d659b9ba1eecc1dd4338e1eec3a/Faculty.jar
21/08/10 21:36:39 WARN manager.MySQLManager: It looks like you are importing from mysql.
21/08/10 21:36:39 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
21/08/10 21:36:39 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
21/08/10 21:36:39 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
21/08/10 21:36:39 INFO mapreduce.ImportJobBase: Beginning import of Faculty
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common
lib/hadoop-auth-2.7.7.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
21/08/10 21:36:40 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
21/08/10 21:36:40 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
21/08/10 21:36:40 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
21/08/10 21:36:40 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
                Total committed heap usage (bytes)=132120576
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=137
21/08/10 21:36:42 INFO mapred.LocalJobRunner: Finishing task: attempt_local1884076220_0001_m_000000_0
21/08/10 21:36:42 INFO mapred.LocalJobRunner: map task executor complete.
21/08/10 21:36:42 INFO mapreduce.Job: Job job_local1884076220_0001 running in uber mode : false
21/08/10 21:36:42 INFO mapreduce.Job:  map 100% reduce 0%
21/08/10 21:36:42 INFO mapreduce.Job: Job job_local1884076220_0001 completed successfully
21/08/10 21:36:42 INFO mapreduce.Job: Counters: 20
        File System Counters
                FILE: Number of bytes read=19489276
                FILE: Number of bytes written=19950174
                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=137
                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=7
                Map output records=7
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=3
                Total committed heap usage (bytes)=132120576
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=137
21/08/10 21:36:42 INFO mapreduce.ImportJobBase: Transferred 137 bytes in 2.0881 seconds (65.6094 bytes/sec)
21/08/10 21:36:42 INFO mapreduce.ImportJobBase: Retrieved 7 records.
```



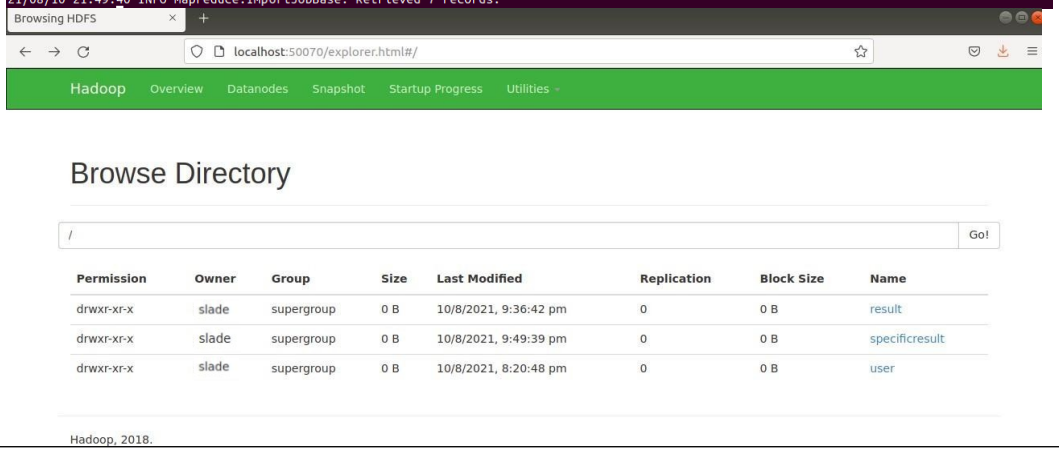


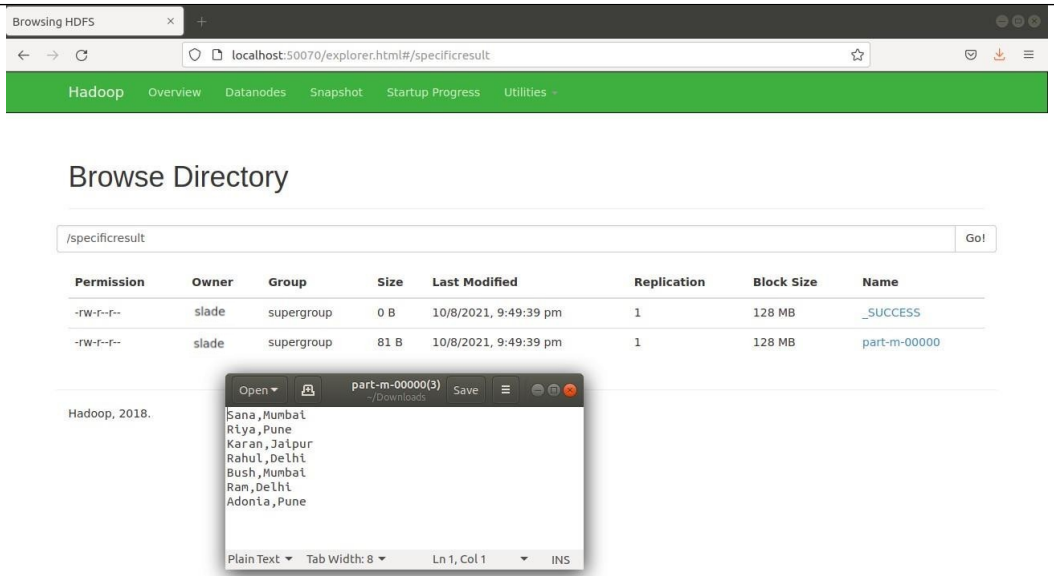
IMPORTING ONLY A SUBSET OF DATA:

2. We will now import only a part of the table Faculty:

```
Warning: /usr/lib/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/lib/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
21/08/10 21:49:36 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
21/08/10 21:49:36 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
21/08/10 21:49:36 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
21/08/10 21:49:36 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 reg
istered via the SPI and manual loading of the driver class is generally unnecessary.
21/08/10 21:49:36 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Faculty` AS t LIMIT 1
21/08/10 21:49:36 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Faculty` AS t LIMIT 1
21/08/10 21:49:36 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-ado/compile/ea8a53962c117324c868c69249e61266/Faculty.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
21/08/10 21:49:37 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ado/compile/ea8a53962c117324c868c69249e61266/Faculty.jar
21/08/10 21:49:37 WARN manager.MySQLManager: It looks like you are importing from mysql.
21/08/10 21:49:37 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
21/08/10 21:49:37 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
21/08/10 21:49:37 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
21/08/10 21:49:37 INFO mapreduce.ImportJobBase: Beginning import of Faculty
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/
lib/hadoop-auth-2.7.7.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
21/08/10 21:49:38 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
21/08/10 21:49:38 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
21/08/10 21:49:38 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
21/08/10 21:49:38 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
21/08/10 21:49:38 INFO db.DBInputFormat: Using read commited transaction isolation
                Total committed heap usage (bytes)=148897792
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=81
21/08/10 21:49:39 INFO mapred.LocalJobRunner: Finishing task: attempt_local1279582788_0001_m_000000_0
21/08/10 21:49:39 INFO mapred.LocalJobRunner: map task executor complete.
21/08/10 21:49:40 INFO mapreduce.Job: Job job_local1279582788_0001 running in uber mode : false
21/08/10 21:49:40 INFO mapreduce.Job:  map 100% reduce 0%
21/08/10 21:49:40 INFO mapreduce.Job: Job job_local1279582788_0001 completed successfully
21/08/10 21:49:40 INFO mapreduce.Job: Counters: 20
        File System Counters
                FILE: Number of bytes read=19488346
                FILE: Number of bytes written=19949224
                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=81
                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=7
                Map output records=7
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=148897792
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=81
21/08/10 21:49:40 INFO mapreduce.ImportJobBase: Transferred 81 bytes in 2.062 seconds (39.2817 bytes/sec)
21/08/10 21:49:40 INFO mapreduce.ImportJobBase: Retrieved 7 records.
```

Text Editor





PROTECTIN G YOUR PASSWORD:

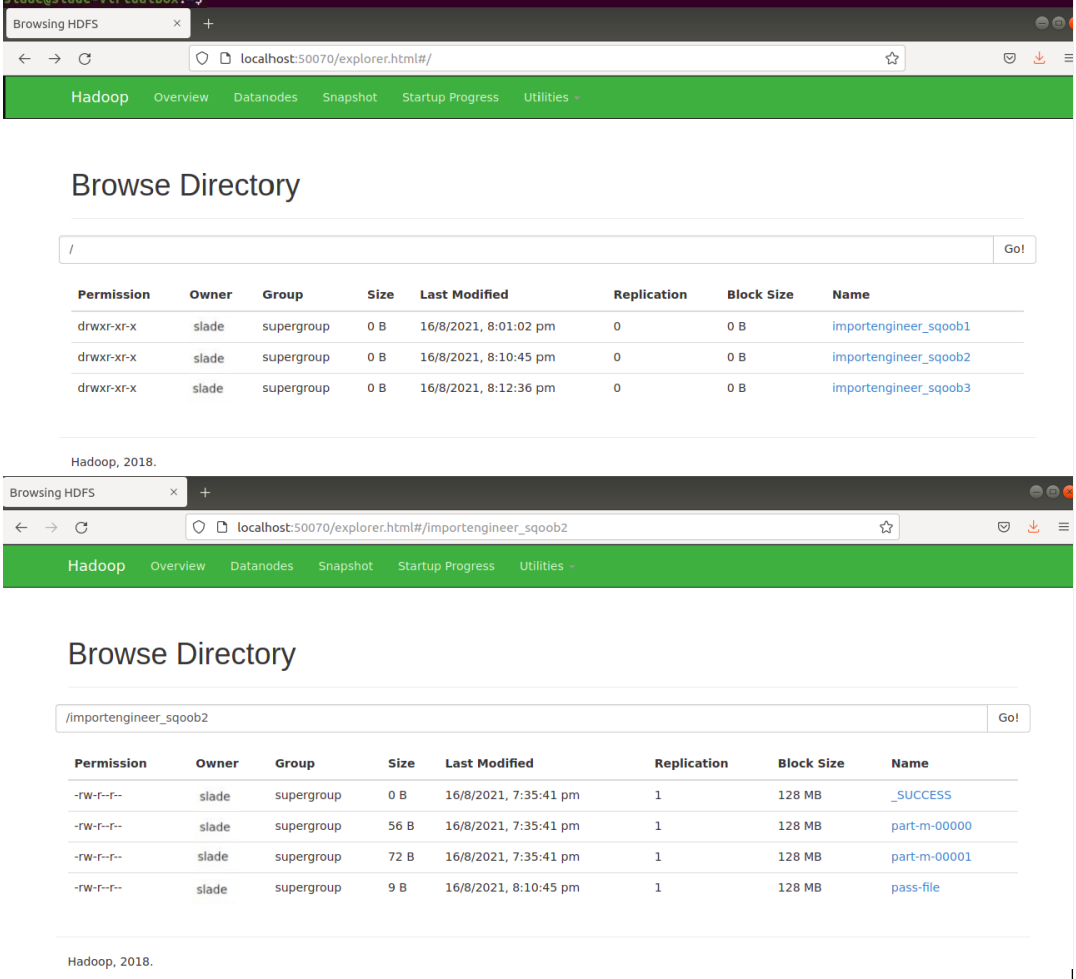
4.There are two ways of specifying the password. First is to write –P in thecommand and then specify the password later on.

```
ado@Toshiba:~$ sqoop import --connect jdbc:mysql://127.0.0.1:3306/engineer --username root -P --table engg_student  --target-dir /importengine
er_sqoob2 -m 2
Warning: /usr/lib/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/lib/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
21/08/16 19:34:52 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
Enter password:
21/08/16 19:35:37 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
21/08/16 19:35:37 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 reg
istered via the SPI and manual loading of the driver class is generally unnecessary.
21/08/16 19:35:37 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `engg_student` AS t LIMIT 1
21/08/16 19:35:37 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `engg_student` AS t LIMIT 1
21/08/16 19:35:37 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-ado/compile/aeab997d316245f8113f1a4eab4d3b28/engg_student.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
21/08/16 19:35:38 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ado/compile/aeab997d316245f8113f1a4eab4d3b28/engg_student.jar
21/08/16 19:35:39 WARN manager.MySQLManager: It looks like you are importing from mysql.
21/08/16 19:35:39 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
21/08/16 19:35:39 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
21/08/16 19:35:39 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
21/08/16 19:35:39 INFO mapreduce.ImportJobBase: Beginning import of engg_student
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/
lib/hadoop-auth-2.7.7.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
21/08/16 19:35:39 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
21/08/16 19:35:39 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
21/08/16 19:35:39 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
21/08/16 19:35:39 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
21/08/16 19:35:39 INFO db.DBInputFormat: Using read commited transaction isolation
                Bytes Written=72
21/08/16 19:35:41 INFO mapred.LocalJobRunner: Finishing task: attempt_local1443023854_0001_m_000001_0
21/08/16 19:35:41 INFO mapred.LocalJobRunner: map task executor complete.
21/08/16 19:35:41 INFO mapreduce.Job: Job job_local1443023854_0001 running in uber mode : false
21/08/16 19:35:41 INFO mapreduce.Job:  map 100% reduce 0%
21/08/16 19:35:41 INFO mapreduce.Job: Job job_local1443023854_0001 completed successfully
21/08/16 19:35:41 INFO mapreduce.Job: Counters: 20
        File System Counters
                FILE: Number of bytes read=38978756
                FILE: Number of bytes written=39900500
                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=184
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
        Map-Reduce Framework
                Map input records=7
                Map output records=7
                Input split bytes=221
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=264241152
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=128
21/08/16 19:35:41 INFO mapreduce.ImportJobBase: Transferred 184 bytes in 2.1112 seconds (87.1531 bytes/sec)
21/08/16 19:35:41 INFO mapreduce.ImportJobBase: Retrieved 7 records.
```

Implement the Second way of specifying the password is writing thepassword in a file and then specifying the file in the command.

```
ado@Toshiba:~$ hadoop fs -put pass-file /importengineer_sqoob2/pass-file
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/
lib/hadoop-auth-2.7.7.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

ado@Toshiba:~$ sqoop import --connect jdbc:mysql://127.0.0.1:3306/engineer --username root -password-file /importengineer_sqoob2/pass-file --t
able engg_student  --target-dir /importengineer_sqoob3 -m 1
Warning: /usr/lib/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/lib/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
21/08/16 20:12:32 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/
lib/hadoop-auth-2.7.7.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
21/08/16 20:12:33 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
21/08/16 20:12:33 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 reg
istered via the SPI and manual loading of the driver class is generally unnecessary.
21/08/16 20:12:33 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `engg_student` AS t LIMIT 1
21/08/16 20:12:33 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `engg_student` AS t LIMIT 1
21/08/16 20:12:33 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-ado/compile/c9129f6c67d1dddcc20b54ff8a0cf3ea/engg_student.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
21/08/16 20:12:34 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ado/compile/c9129f6c67d1dddcc20b54ff8a0cf3ea/engg_student.jar
21/08/16 20:12:34 WARN manager.MySQLManager: It looks like you are importing from mysql.
21/08/16 20:12:34 WARN manager.MySQLManager: This transfer can be faster! Use the --dlrect
21/08/16 20:12:34 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
21/08/16 20:12:34 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
21/08/16 20:12:34 INFO mapreduce.ImportJobBase: Beginning import of engg_student
21/08/16 20:12:34 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
21/08/16 20:12:34 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
21/08/16 20:12:34 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
21/08/16 20:12:34 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
21/08/16 20:12:34 INFO db.DBInputFormat: Using read commited transaction isolation
```

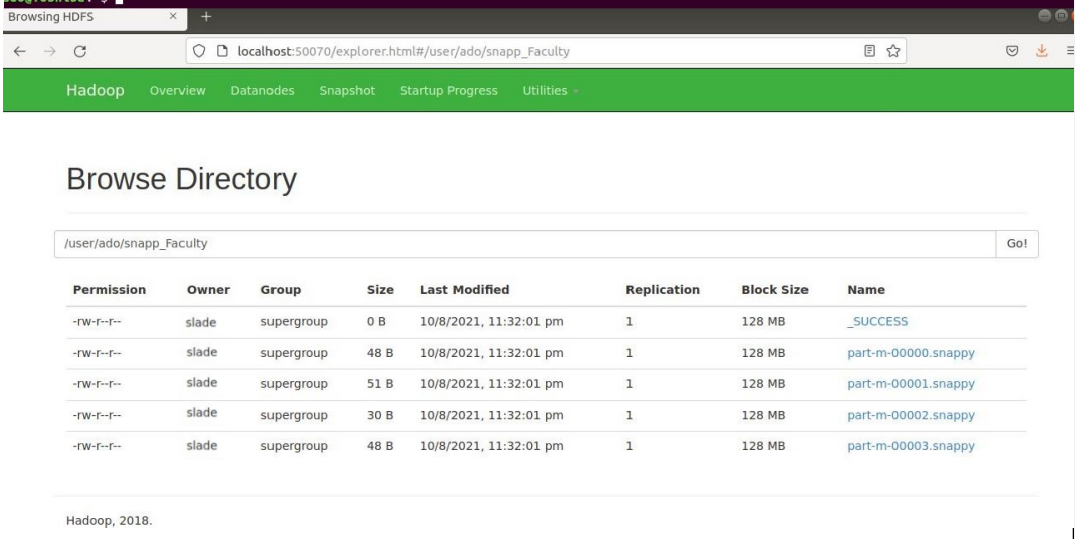
```
                Total committed heap usage (bytes)=132120576
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=128
21/08/16 20:12:36 INFO mapred.LocalJobRunner: Finishing task: attempt_local996615860_0001_m_000000_0
21/08/16 20:12:36 INFO mapred.LocalJobRunner: map task executor complete.
21/08/16 20:12:36 INFO mapreduce.Job: Job job_local996615860_0001 running in uber mode : false
21/08/16 20:12:36 INFO mapreduce.Job:  map 100% reduce 0%
21/08/16 20:12:36 INFO mapreduce.Job: Job job_local996615860_0001 completed successfully
21/08/16 20:12:36 INFO mapreduce.Job: Counters: 20
        File System Counters
                FILE: Number of bytes read=19489121
                FILE: Number of bytes written=19948517
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=9
                HDFS: Number of bytes written=128
                HDFS: Number of read operations=7
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
        Map-Reduce Framework
                Map input records=7
                Map output records=7
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=3
                Total committed heap usage (bytes)=132120576
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=128
21/08/16 20:12:36 INFO mapreduce.ImportJobBase: Transferred 128 bytes in 2.0928 seconds (61.1623 bytes/sec)
21/08/16 20:12:36 INFO mapreduce.ImportJobBase: Retrieved 7 records.
slade@slade-VirtualBox:~$
```



COMPRESSING IMPORTED DATA:

4. Use command to compress the imported data.

```
ado@Toshiba:~$ sqoop import --connect jdbc:mysql://127.0.0.1:3306/emp --username root --password Adonia@12 --table Faculty --target-dir /user/
ado/snapp_Faculty --compression-codec org.apache.hadoop.io.compress.SnappyCodec
Warning: /usr/lib/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/lib/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
21/08/10 23:31:57 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
21/08/10 23:31:57 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
21/08/10 23:31:57 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
21/08/10 23:31:57 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 reg
istered via the SPI and manual loading of the driver class is generally unnecessary.
21/08/10 23:31:57 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Faculty` AS t LIMIT 1
21/08/10 23:31:57 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `Faculty` AS t LIMIT 1
21/08/10 23:31:57 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-ado/compile/a2aeda7e422d39a463c26b59cd1c93d8/Faculty.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
21/08/10 23:31:59 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ado/compile/a2aeda7e422d39a463c26b59cd1c93d8/Faculty.jar
21/08/10 23:31:59 WARN manager.MySQLManager: It looks like you are importing from mysql.
21/08/10 23:31:59 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
21/08/10 23:31:59 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
21/08/10 23:31:59 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
21/08/10 23:31:59 INFO mapreduce.ImportJobBase: Beginning import of Faculty
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/
lib/hadoop-auth-2.7.7.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
21/08/10 23:31:59 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
21/08/10 23:31:59 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
21/08/10 23:31:59 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
21/08/10 23:31:59 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
21/08/10 23:32:00 INFO db.DBInputFormat: Using read commited transaction isolation
                Total committed heap usage (bytes)=132120576
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=48
21/08/10 23:32:01 INFO mapred.LocalJobRunner: Finishing task: attempt_local3483673_0001_m_000003_0
21/08/10 23:32:01 INFO mapred.LocalJobRunner: map task executor complete.
21/08/10 23:32:01 INFO mapreduce.Job: Job job_local3483673_0001 running in uber mode : false
21/08/10 23:32:01 INFO mapreduce.Job:  map 100% reduce 0%
21/08/10 23:32:01 INFO mapreduce.Job: Job job_local3483673_0001 completed successfully
21/08/10 23:32:01 INFO mapreduce.Job: Counters: 20
        File System Counters
                FILE: Number of bytes read=77960848
                FILE: Number of bytes written=79782696
                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=453
                HDFS: Number of read operations=34
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=24
        Map-Reduce Framework
                Map input records=7
                Map output records=7
                Input split bytes=393
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=2
                Total committed heap usage (bytes)=528482304
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=177
21/08/10 23:32:01 INFO mapreduce.ImportJobBase: Transferred 453 bytes in 2.0897 seconds (216.7817 bytes/sec)
21/08/10 23:32:01 INFO mapreduce.ImportJobBase: Retrieved 7 records.
ado@Toshiba:~$
```



mmm

Exporting DATA FROM HDFS to RELATIONAL DATABASE:

5. Export the data

Create database engineer

```
ado@Toshiba:~$ mysql -u root -p
Enter password:
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 25
Server version: 5.7.35-0ubuntu0.18.04.1 (Ubuntu)

Copyright (c) 2000, 2021, Oracle and/or its affiliates.

Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.

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

mysql> create database engineer
    -> ;
Query OK, 1 row affected (0.01 sec)

mysql> create table student;
ERROR 1046 (3D000): No database selected
mysql> use engineer;
Database changed
```

```
mysql> use engineer;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> show tables;
+--------------------+
| Tables_in_engineer |
+--------------------+
| engg_student       |
+--------------------+
1 row in set (0.00 sec)

mysql> select * from engg_student;
+---------+--------+--------+-------+
| engg_id | f_name | l_name | dept  |
+---------+--------+--------+-------+
|      51 | Adonia | Seq    | COMPS |
|      52 | Jaden  | Smith  | MECH  |
|      53 | Jenny  | Vaz    | IT    |
|      54 | John   | Aly    | EXTC  |
|      55 | Ann    | Mary   | COMPS |
|      56 | Tom    | Thomas | IT    |
|      57 | Nicci  | Ferns  | EXTC  |
+---------+--------+--------+-------+
7 rows in set (0.05 sec)
```

```
ado@Toshiba:~$ sqoop import --connect jdbc:mysql://127.0.0.1:3306/engineer --username root --password Adonia@12 --table engg_student --target
dir /importengineer_sqoob -m 2
Warning: /usr/lib/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/lib/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
21/08/11 23:20:25 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
21/08/11 23:20:25 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
21/08/11 23:20:26 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
21/08/11 23:20:26 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 re
istered via the SPI and manual loading of the driver class is generally unnecessary.
21/08/11 23:20:26 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `engg_student` AS t LIMIT 1
21/08/11 23:20:26 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `engg_student` AS t LIMIT 1
21/08/11 23:20:26 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-ado/compile/e0f46311396b8ee3fca5ead35fbdbbd6/engg_student.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
21/08/11 23:20:27 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ado/compile/e0f46311396b8ee3fca5ead35fbdbbd6/engg_student.jar
21/08/11 23:20:27 WARN manager.MySQLManager: It looks like you are importing from mysql.
21/08/11 23:20:27 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
21/08/11 23:20:27 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
21/08/11 23:20:27 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
21/08/11 23:20:27 INFO mapreduce.ImportJobBase: Beginning import of engg_student
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=72
21/08/11 23:20:29 INFO mapred.LocalJobRunner: Finishing task: attempt_local2074460199_0001_m_000001_0
21/08/11 23:20:29 INFO mapred.LocalJobRunner: map task executor complete.
21/08/11 23:20:30 INFO mapreduce.Job: Job job_local2074460199_0001 running in uber mode : false
21/08/11 23:20:30 INFO mapreduce.Job:  map 100% reduce 0%
21/08/11 23:20:30 INFO mapreduce.Job: Job job_local2074460199_0001 completed successfully
21/08/11 23:20:30 INFO mapreduce.Job: Counters: 20
        File System Counters
                FILE: Number of bytes read=38978756
                FILE: Number of bytes written=39900496
                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=184
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
        Map-Reduce Framework
                Map input records=7
                Map output records=7
                Input split bytes=221
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=264241152
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=128
21/08/11 23:20:30 INFO mapreduce.ImportJobBase: Transferred 184 bytes in 2.0544 seconds (89.5641 bytes/sec)
21/08/11 23:20:30 INFO mapreduce.ImportJobBase: Retrieved 7 records.
```

Check table contents

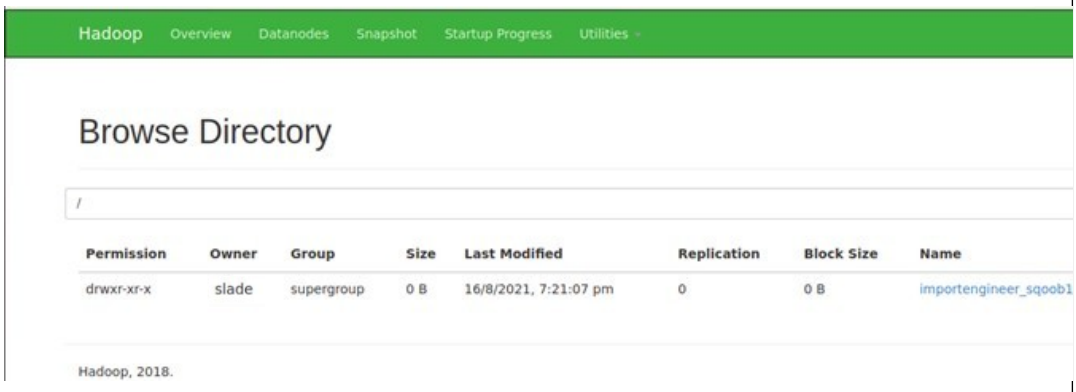
```
ado@Toshiba:~$ hadoop fs -cat /importengineer_sqoob/part-m-00000
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/
lib/hadoop-auth-2.7.7.jar) to 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
51,Adonia,Seq,COMPS
52,Jaden,Smith,MECH
53,Jenny,Vaz,IT
ado@Toshiba:~$ hadoop fs -cat /importengineer_sqoob/part-m-00001
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/
lib/hadoop-auth-2.7.7.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
54,John,Aly,EXTC
55,Ann,Mary,COMPS
56,Tom,Thomas,IT
57,Nicci,Ferns,EXTC
```

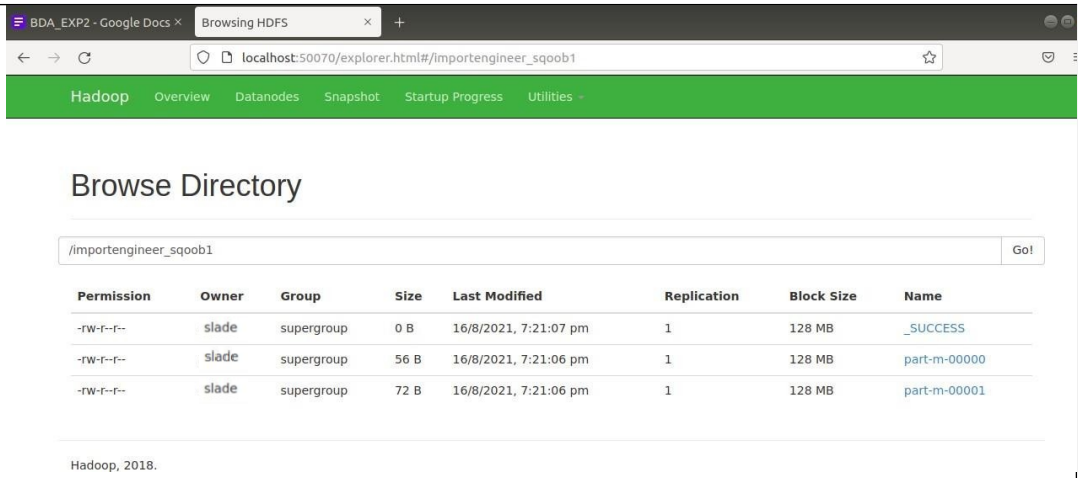
Create new table for export

```
mysql> desc engg_student;
+---------+-------------+------+-----+---------+-------+
| Field   | Type        | Null | Key | Default | Extra |
+---------+-------------+------+-----+---------+-------+
| engg_id | int(11)     | NO   | PRI | NULL    |       |
| f_name  | varchar(30) | NO   |     | NULL    |       |
| l_name  | varchar(30) | NO   |     | NULL    |       |
| dept    | varchar(20) | NO   |     | NULL    |       |
+---------+-------------+------+-----+---------+-------+
4 rows in set (0.00 sec)

mysql> create table expo_engg(engg_id int NOT NULL,f_name varchar(30)
NOT NULL, l_name varchar(30) NOT NULL, dept varchar(20) NOT NULL, PRIM
ARY KEY(engg_id));
Query OK, 0 rows affected (0.35 sec)

mysql> desc expo_engg;
+---------+-------------+------+-----+---------+-------+
| Field   | Type        | Null | Key | Default | Extra |
+---------+-------------+------+-----+---------+-------+
| engg_id | int(11)     | NO   | PRI | NULL    |       |
| f_name  | varchar(30) | NO   |     | NULL    |       |
| l_name  | varchar(30) | NO   |     | NULL    |       |
| dept    | varchar(20) | NO   |     | NULL    |       |
+---------+-------------+------+-----+---------+-------+
4 rows in set (0.00 sec)

mysql> select * from expo_engg;
Empty set (0.00 sec)
```

```
ado@Toshiba:~$ sqoop import --connect jdbc:mysql://127.0.0.1:3306/engineer --username root --password Adonia@12 --table engg_st
-dir /importengineer_sqoob1 -m 2
Warning: /usr/lib/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/lib/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
21/08/16 19:21:02 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
21/08/16 19:21:02 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
21/08/16 19:21:02 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
21/08/16 19:21:02 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 au
tomatically registered via the SPI and manual loading of the driver class is generally unnecessary.
21/08/16 19:21:03 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `engg_student` AS t LIMIT 1
21/08/16 19:21:03 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `engg_student` AS t LIMIT 1
21/08/16 19:21:03 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-ado/compile/d17765b299c9cc2b1e95f13540044df9/engg_student.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
21/08/16 19:21:04 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ado/compile/d17765b299c9cc2b1e95f13540044df9/engg_s
21/08/16 19:21:04 WARN manager.MySQLManager: It looks like you are importing from mysql.
21/08/16 19:21:04 WARN manager.MySQLManager: This transfer can be faster! Use the --direct
21/08/16 19:21:04 WARN manager.MySQLManager: option to exercise a MySQL-specific fast path.
21/08/16 19:21:04 INFO manager.MySQLManager: Setting zero DATETIME behavior to convertToNull (mysql)
21/08/16 19:21:04 INFO mapreduce.ImportJobBase: Beginning import of engg_student
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share
lib/hadoop-auth-2.7.7.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
21/08/16 19:21:04 INFO Configuration.deprecation: mapred.jar                nstead, use mapreduce.job.jar
21/08/16 19:21:04 INFO Configuration.deprecation: mapred.job.               cated. Instead, use mapreduce.jobtracker.address
21/08/16 19:21:04 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
21/08/16 19:21:04 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
                Total committed heap usage (bytes)=132120576
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=72
21/08/16 19:21:07 INFO mapred.LocalJobRunner: Finishing task: attempt_local1457900323_0001_m_000001_0
21/08/16 19:21:07 INFO mapred.LocalJobRunner: map task executor complete.
21/08/16 19:21:07 INFO mapreduce.Job: Job job_local1457900323_0001 running in uber mode : false
21/08/16 19:21:07 INFO mapreduce.Job:  map 100% reduce 0%
21/08/16 19:21:08 INFO mapreduce.Job: Job job_local1457900323_0001 completed successfully
21/08/16 19:21:08 INFO mapreduce.Job: Counters: 20
        File System Counters
                FILE: Number of bytes read=38978756
                FILE: Number of bytes written=39900500
                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=184
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=8
        Map-Reduce Framework
                Map input records=7
                Map output records=7
                Input split bytes=221
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=264241152
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=128
21/08/16 19:21:08 INFO mapreduce.ImportJobBase: Transferred 184 bytes in 3.0899 seconds (59.5483 bytes/sec)
21/08/16 19:21:08 INFO mapreduce.ImportJobBase: Retrieved 7 records.
```

Video player






```
ado@Toshiba:~$ hadoop fs -cat /importengineer_sqoob1/part-m-00000
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/
lib/hadoop-auth-2.7.7.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
51,Adonia,Seq,COMPS
52,Jaden,Smith,MECH
53,Jenny,Vaz,IT
ado@Toshiba:~$ hadoop fs -cat /importengineer_sqoob1/part-m-00001
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/
lib/hadoop-auth-2.7.7.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
54,John,Aly,EXTC
55,Ann,Mary,COMPS
56,Tom,Thomas,IT
57,Nicci,Ferns,EXTC
ado@Toshiba:~$ sqoop export --connect jdbc:mysql://127.0.0.1:3306/engineer --username root --password Adonia@12 --table expo_engg --export-di
 /importengineer_sqoob1
Warning: /usr/lib/sqoop/../hbase does not exist! HBase imports will fail.
Please set $HBASE_HOME to the root of your HBase installation.
Warning: /usr/lib/sqoop/../hcatalog does not exist! HCatalog jobs will fail.
Please set $HCAT_HOME to the root of your HCatalog installation.
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
Warning: /usr/lib/sqoop/../zookeeper does not exist! Accumulo imports will fail.
Please set $ZOOKEEPER_HOME to the root of your Zookeeper installation.
21/08/16 19:25:28 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6
21/08/16 19:25:28 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
21/08/16 19:25:28 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
21/08/16 19:25:28 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 re
istered via the SPI and manual loading of the driver class is generally unnecessary.
21/08/16 19:25:29 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `expo_engg` AS t LIMIT 1
21/08/16 19:25:29 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM `expo_engg` AS t LIMIT 1
21/08/16 19:25:29 INFO orm.CompilationManager: HADOOP_MAPRED_HOME is /usr/local/hadoop
Note: /tmp/sqoop-ado/compile/c267a4913bf8a97ab29e8eb46993f7d1/expo_engg.java uses or overrides a deprecated API.
Note: Recompile with -Xlint:deprecation for details.
21/08/16 19:25:30 INFO orm.CompilationManager: Writing jar file: /tmp/sqoop-ado/compile/c267a4913bf8a97ab29e8eb46993f7d1/expo_engg.jar
21/08/16 19:25:30 INFO mapreduce.ExportJobBase: Beginning export of expo_engg
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common
lib/hadoop-auth-2.7.7.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
21/08/16 19:25:30 INFO Configuration.deprecation: mapred.jar is deprecated. Instead, use mapreduce.job.jar
21/08/16 19:25:31 INFO Configuration.deprecation: mapred.reduce.tasks.speculative.execution is deprecated. Instead, use mapreduce.reduce.spec
lative
21/08/16 19:25:31 INFO Configuration.deprecation: mapred.map.tasks.speculative.execution is deprecated. Instead, use mapreduce.map.speculativ
21/08/16 19:25:31 INFO Configuration.deprecation: mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.address
21/08/16 19:25:31 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
21/08/16 19:25:31 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
21/08/16 19:25:31 INFO input.FileInputFormat: Total input paths to process : 2
21/08/16 19:25:31 INFO input.FileInputFormat: Total input paths to process : 2
                Total committed heap usage (bytes)=148897792
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
21/08/16 19:25:32 INFO mapred.LocalJobRunner: Finishing task: attempt_local522259445_0001_m_000002_0
21/08/16 19:25:32 INFO mapred.LocalJobRunner: map task executor complete.
21/08/16 19:25:33 INFO mapreduce.Job: Job job_local522259445_0001 running in uber mode : false
21/08/16 19:25:33 INFO mapreduce.Job:  map 100% reduce 0%
21/08/16 19:25:33 INFO mapreduce.Job: Job job_local522259445_0001 completed successfully
21/08/16 19:25:33 INFO mapreduce.Job: Counters: 20
        File System Counters
                FILE: Number of bytes read=58470603
                FILE: Number of bytes written=59844012
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=486
                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=557
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=446693376
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=0
21/08/16 19:25:33 INFO mapreduce.ExportJobBase: Transferred 486 bytes in 2.0199 seconds (240.6027 bytes/sec)
21/08/16 19:25:33 INFO mapreduce.ExportJobBase: Exported 7 records.
ado@Toshiba:~$
```

```
mysql> show tables;
+--------------------+
| Tables_in_engineer |
+--------------------+
| engg_student       |
| expo_engg          |
+--------------------+
2 rows in set (0.00 sec)

mysql> select * from engg_student;
+---------+--------+--------+-------+
| engg_id | f_name | l_name | dept  |
+---------+--------+--------+-------+
|      51 | Adonia | Seq    | COMPS |
|      52 | Jaden  | Smith  | MECH  |
|      53 | Jenny  | Vaz    | IT    |
|      54 | John   | Aly    | EXTC  |
|      55 | Ann    | Mary   | COMPS |
|      56 | Tom    | Thomas | IT    |
|      57 | Nicci  | Ferns  | EXTC  |
+---------+--------+--------+-------+
7 rows in set (0.00 sec)

mysql> select * from expo_engg;;
+---------+--------+--------+-------+
| engg_id | f_name | l_name | dept  |
+---------+--------+--------+-------+
|      51 | Adonia | Seq    | COMPS |
|      52 | Jaden  | Smith  | MECH  |
|      53 | Jenny  | Vaz    | IT    |
|      54 | John   | Aly    | EXTC  |
|      55 | Ann    | Mary   | COMPS |
|      56 | Tom    | Thomas | IT    |
|      57 | Nicci  | Ferns  | EXTC  |
+---------+--------+--------+-------+
7 rows in set (0.00 sec)
```

Conclusion:	Students will be able to use Sqoop tool for transferring data between Hadoop &relational databases
References:	http://moodle.dbit.in/ https://www.edureka.co/blog/apache-sqoop-tutorial/ https://dwgeek.com/sqoop-command-with-secure-password.html/

Experiment No: 3

Name: Alston Fernandes **Roll No: 19**

Topic:	Implementing distinct word count using MapReduce.
Prerequisite:	Basic Java programming, Hadoop and MapReduce knowledge is required
Mapping With COs:	CSL702.2
Objective:	To write the program for mapper class, reducer class and driver class to find the number of distinct words present in the input file.
Outcomes:	Students will be able to write the program for mapper class, reducer class and driver class to find the number of distinct words present in the input file and be able to produce part-r file as output.
Instructions:	This experiment is a compulsory experiment. All the students are required to perform this experiment individually.

Deliverables:	1. Give the details of the input file for this program. **input.txt** dog cat deer car cat car dog boy dog car welcome all all is well well done cat deer deer is car cat dog car done cat is dog well deer dog cat deer car cat car dog boy dog car welcome all all is well well done cat deer deer is car cat dog car done cat is dog well deer dog cat deer

	car cat car dog boy dog car welcome all all is well well done cat deer deer is car cat dog car done cat is dog well deer dog cat deer car cat car dog boy dog car welcome all all is well well done cat deer deer is car cat dog car done cat is dog well deer dog cat deer car cat car dog boy dog car

```
welcome all
all is well
well
done
cat deer
deer is car
cat dog
car done
cat is dog
well deer
dog cat
deer
car
cat car
dog
boy dog car
welcome all
all is well
well
done
cat deer
deer is car
cat dog
car done
cat is dog
well deer
```

2. Write the code of mapper, reducer and driver.

```
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;
import org.apache.hadoop.util.GenericOptionsParser;


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();
      String[] otherArgs = new GenericOptionsParser(conf,
args).getRemainingArgs();
      if (otherArgs.length < 2) {
      System.err.println("Usage: wordcount <in> [<in>...]
<out>");
      System.exit(2);
      }
      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);
      for (int i = 0; i < otherArgs.length - 1; ++i) {
      FileInputFormat.addInputPath(job, new
Path(otherArgs[i]));
      }
      FileOutputFormat.setOutputPath(job,
      new Path(otherArgs[otherArgs.length - 1]));
      System.exit(job.waitForCompletion(true) ? 0 : 1);
  }
}
```

3. Output for this program.(Snapshots)

Starting hadoop

```
dbit@complab3:~$ hdfs namenode -format
23/09/10 12:17:03 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = complab3/127.0.1.1
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 2.7.7
STARTUP_MSG:   classpath = /usr/local/hadoop/etc/hadoop:/usr/local/hadoop/share/hadoop/common/lib/sl
share/hadoop/common/lib/hamcrest-core-1.3.jar:/usr/local/hadoop/share/hadoop/common/lib/jackson-mapp
/hadoop/common/lib/log4j-1.2.17.jar:/usr/local/hadoop/share/hadoop/common/lib/paranamer-2.3.jar:/usr
```

```
dbit@complab3:~$ start-all.sh
This script is Deprecated. Instead use start-dfs.sh and start-yarn.sh
Starting namenodes on [localhost]
dbit@localhost's password:
localhost: starting namenode, logging to /usr/local/hadoop/logs/hadoop-dbit-namenode-complab3.out
dbit@localhost's password:
localhost: starting datanode, logging to /usr/local/hadoop/logs/hadoop-dbit-datanode-complab3.out
Starting secondary namenodes [0.0.0.0]
dbit@0.0.0.0's password:
0.0.0.0: starting secondarynamenode, logging to /usr/local/hadoop/logs/hadoop-dbit-secondarynamenode-complab3.out
starting yarn daemons
starting resourcemanager, logging to /usr/local/hadoop/logs/yarn-dbit-resourcemanager-complab3.out
dbit@localhost's password:
localhost: starting nodemanager, logging to /usr/local/hadoop/logs/yarn-dbit-nodemanager-complab3.out
dbit@complab3:~$
```

Put file into HDFS

```
dbit@complab3:~$ hadoop fs -mkdir /WordCountExp
dbit@complab3:~$ hadoop fs -put Desktop/WordCount/input.txt /WordCountExp
dbit@complab3:~$
```

Compile file

```
dbit@complab3:~$ javac -classpath ${HADOOP_CLASSPATH} -d Desktop/WordCount/classes/ Desktop/WordCount/WordCount.java
dbit@complab3:~$ cd Desktop/WordCount/
dbit@complab3:~/Desktop/WordCount$ jar -cvf exp.jar classes/
added manifest
adding: classes/(in = 0) (out= 0)(stored 0%)
adding: classes/WordCount.class(in = 1927) (out= 1052)(deflated 45%)
adding: classes/WordCount$TokenizerMapper.class(in = 1752) (out= 764)(deflated 56%)
adding: classes/WordCount$IntSumReducer.class(in = 1755) (out= 750)(deflated 57%)
dbit@complab3:~/Desktop/WordCount$
```

Pass the file to hadoop MapReducer

```
dbit@complab3:~/Desktop/WordCount$ hadoop jar exp.jar WordCount /WordCountExp /WordCountExp/Out
23/09/10 20:18:21 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/09/10 20:18:23 INFO input.FileInputFormat: Total input paths to process : 1
23/09/10 20:18:23 INFO mapreduce.JobSubmitter: number of splits:1
23/09/10 20:18:23 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1694356176649_0001
23/09/10 20:18:29 INFO impl.YarnClientImpl: Submitted application application_1694356176649_0001
23/09/10 20:18:29 INFO mapreduce.Job: The url to track the job: http://complab3:8088/proxy/application_1694356176649_0001/
23/09/10 20:18:29 INFO mapreduce.Job: Running job: job_1694356176649_0001
23/09/10 20:18:52 INFO mapreduce.Job: Job job_1694356176649_0001 running in uber mode : false
23/09/10 20:18:52 INFO mapreduce.Job:  map 0% reduce 0%
23/09/10 20:18:59 INFO mapreduce.Job:  map 100% reduce 0%
```

```
        Data-local map tasks=1
        Total time spent by all maps in occupied slots (ms)=4991
        Total time spent by all reduces in occupied slots (ms)=4382
        Total time spent by all map tasks (ms)=4991
        Total time spent by all reduce tasks (ms)=4382
        Total vcore-milliseconds taken by all map tasks=4991
        Total vcore-milliseconds taken by all reduce tasks=4382
        Total megabyte-milliseconds taken by all map tasks=5110784
        Total megabyte-milliseconds taken by all reduce tasks=4487168
    Map-Reduce Framework
        Map input records=16
        Map output records=18
        Map output bytes=153
        Map output materialized bytes=152
        Input split bytes=116
        Combine input records=18
        Combine output records=14
        Reduce input groups=14
        Reduce shuffle bytes=152
        Reduce input records=14
        Reduce output records=14
        Spilled Records=28
        Shuffled Maps =1
        Failed Shuffles=0
        Merged Map outputs=1
        GC time elapsed (ms)=175
        CPU time spent (ms)=1890
        Physical memory (bytes) snapshot=431947776
        Virtual memory (bytes) snapshot=3835953152
        Total committed heap usage (bytes)=295174144
    Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
```

Output

```
dbit@complab3:~/Desktop/WordCount$ hadoop fs -cat /WordCountExp/Output/part-r-00000
dog     30
cat     30
car     30
deer    24
well    18
is      18
done    12
all     12
welcome 6
boy     6
dbit@complab3:~/Desktop/WordCount$
```

Conclusion:	Thus we are able to execute the Mapreduce program to count distinct words in the input file using hadoop pseudo distributed mode.
References:	https://learnomate.org/steps-to-resolve-when-datanode-services-is-not-starting/ https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm

Don Bosco Institute of Technology

Department of Computer Engineering

Academic year – 2022-2023

Big Data Analytics

Assessment Rubric for Experiment No.: 03

Performance Date :
Submission Date :

Title of Experiment : Implementing distinct word count using MapReduce.

Year and Semester : IV^{rh} Year and VIIth Semester
Batch :
Name of Student :
Roll No. :

Performance	**Poor**	**Satisfactory**	**Good**	**Excellent**	
	1 point	2 points	3 points	4 points	
Results and Documentations	**Poor**	**Satisfactory**	**Good**	**Excellent**	
	1 point	2 points	3 points	4 points	
Viva	**Poor**	**Satisfactory**	**Good**	**Excellent**	
	1 point	2 points	3 points	4 points	
Timely Submission	**Submission beyond 7 days of the deadline**	**Late submission till 7 days**	**Submission on time**		
	1 points	2 points	3 points		

Faculty Incharge : Ms. Sana Shaikh

Faculty: Sana Shaikh **Lab Manual - BDA** **2023-2024**

Student Name: Alston Fernandes **Experiment No: 4** **Batch: A**
Roll No.: 19

Topic:	Implement Bloom Filter using any programming language.
Prerequisite:	- Familiarity with the programming language - Basic concept of Bloom Filter
Mapping With COs:	CSL702.5
Objective:	Implement and apply Bloom filters for any one appropriate real world application.
Outcomes:	Students will be able to understand the concept of Bloom filter and its usage and also be able to implement it for any real-world problem.
Instructions:	This experiment is a compulsory experiment. All the students are required to perform this experiment in a group. [same group as for Mini Project]
Deliverables:	**Submission on Moodle:** - **Explain one Real World Filter.** A real-world example of a filter-like library where you can add and search for books is a digital library catalogue or a library management system. These systems are commonly used by libraries and educational institutions to keep track of their book collections and provide users with the ability to find and access books. Here's an explanation of how such a system works: **Digital Library Catalogue or Library Management System** 1. **Book Database:** The system maintains a database that stores information about the books in the library's collection. Each book record typically includes details such as the book's title, author, publication date, ISBN, genre, availability status, and more. 2. **Add Books:** Librarians or authorized personnel can add new books to the catalogue. They input the book's information into the system, and the data is stored in the database. Some systems might also support the automatic addition of books using barcode scanning or ISBN lookup. 3. **Search Functionality:** Users, such as library patrons or staff, can search for books in the catalogue. They can search by various criteria, including title, author,

keyword, ISBN, genre, and more. The system uses search algorithms to quickly locate relevant book records based on the user's query.

4. **Filters and Sorting:** Users can often apply filters and sorting options to refine their search results. For example, they can filter books by genre, availability, or publication date. Sorting options allow users to order the results alphabetically, by relevance, or by other criteria.

5. **Availability Status:** The system keeps track of the availability status of each book. It indicates whether a book is available for borrowing, checked out, on hold, or missing. Users can see this status when searching for books.

6. **User Accounts:** Many library systems require users to create accounts. Registered users can have additional features such as reserving books, viewing their borrowing history, and extending loan periods.

7**. Reservations and Checkouts:** Users can reserve books that are currently checked out by others. When a reserved book becomes available, the system notifies the user, and they can check it out. Users can also check out books directly from the catalogue.

A digital library catalogue or library management system is an example of a real-world filter-like library where users can add and search for books. It serves as a centralized repository of information about a library's book collection and provides features for both library staff and patrons to efficiently manage and access books.

- **INPUT:**

- Librarians can add new books to the catalogue. They input the book's Name as information.
- Librarians can search for books in the catalogue.

OUTPUT:

- Status of Book Whether it's present or not.

- **Bloom Filter CODE:** (language: Python)

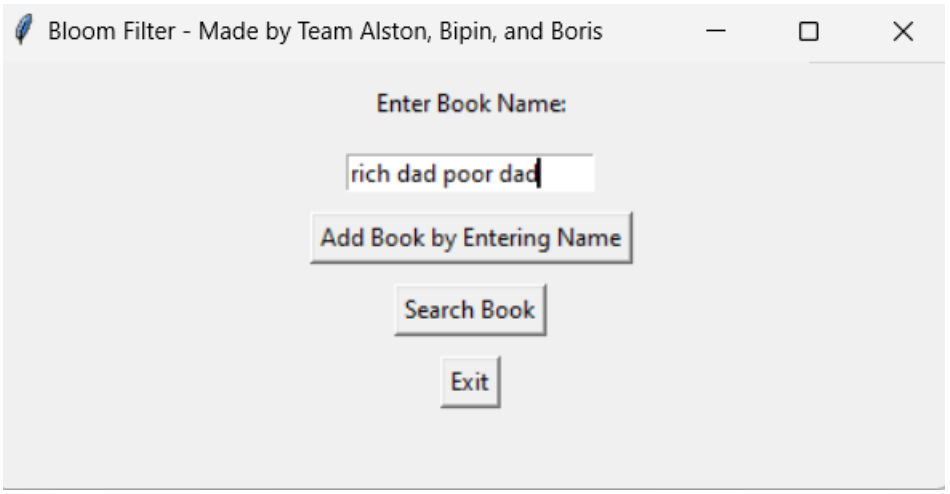
```
import hashlib
import tkinter as tk
from tkinter import messagebox

class BloomFilter:
    def __init__(self, size, hash_functions):
        self.size = size
        self.bit_array = [0] * size
        self.hash_functions = hash_functions
```

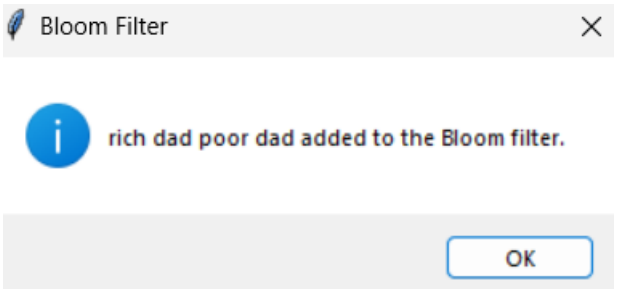
```
    def add(self, element):
        for hash_func in self.hash_functions:
            index = hash_func(element) % self.size
            self.bit_array[index] = 1

    def contains(self, element):
        for hash_func in self.hash_functions:
            index = hash_func(element) % self.size
            if self.bit_array[index] == 0:
                return False
        return True

def add_element():
    element = entry.get()
    if element:
        bloom_filter.add(element)
        entry.delete(0, tk.END)
        messagebox.showinfo("Bloom Filter", f"{element}
added to the Bloom filter.")

def search_element():
    element = entry.get()
    if element:
        if bloom_filter.contains(element):
            messagebox.showinfo("Bloom Filter", f"{element}
may be in the set.")
        else:
            messagebox.showinfo("Bloom Filter", f"{element}
is definitely not in the set.")
        entry.delete(0, tk.END)

def exit_program():
    window.destroy()

size = 20
num_hash_functions = 3

hash_functions = [
    lambda x: int(hashlib.md5(x.encode()).hexdigest(), 16),
    lambda x: int(hashlib.sha1(x.encode()).hexdigest(), 16),
    lambda x: int(hashlib.sha256(x.encode()).hexdigest(),
16)
]

bloom_filter = BloomFilter(size, hash_functions)

window = tk.Tk()
```

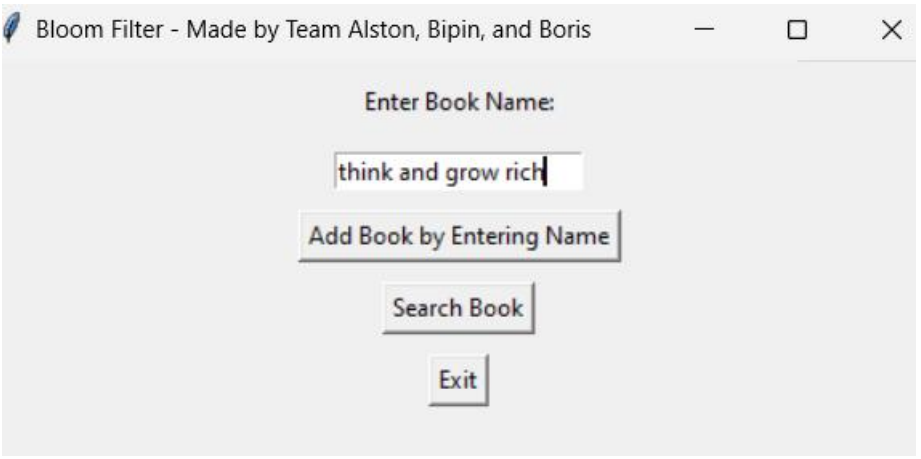
```
window.title("Bloom Filter - Made by Team Alston, Bipin, and
Boris")
window.geometry("400x300")
window.resizable(True, True)

label = tk.Label(window, text="Enter Book Name:")
entry = tk.Entry(window)
add_button = tk.Button(window, text="Add Book by Entering
Name", command=add_element)
search_button = tk.Button(window, text="Search Book",
command=search_element)
exit_button = tk.Button(window, text="Exit",
command=exit_program)

label.pack(pady=10)
entry.pack(pady=5)
add_button.pack(pady=5)
search_button.pack(pady=5)
exit_button.pack(pady=5)

window.mainloop()
```

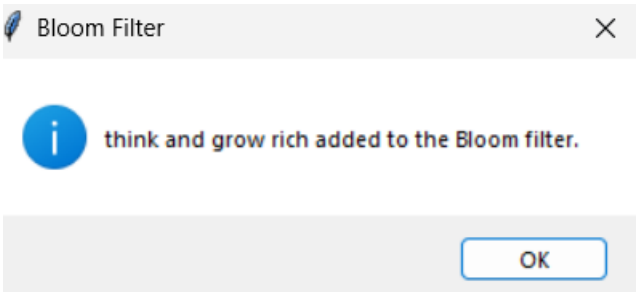
- Results: Input and Output

Adding book

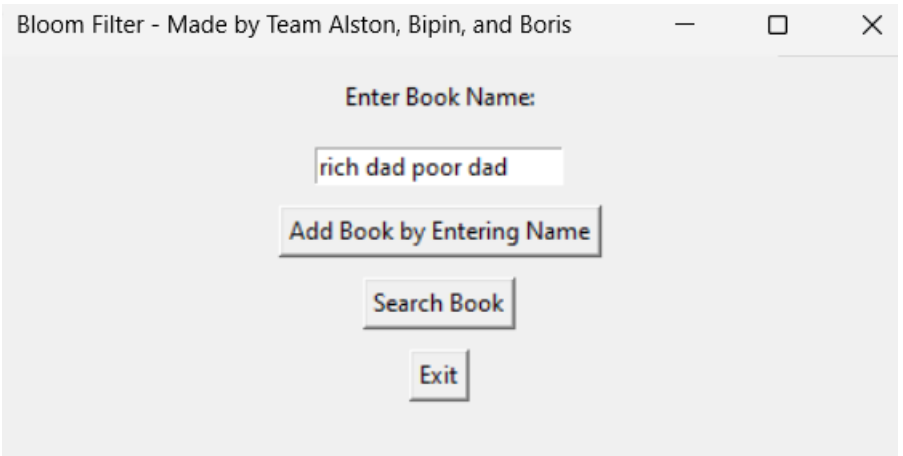


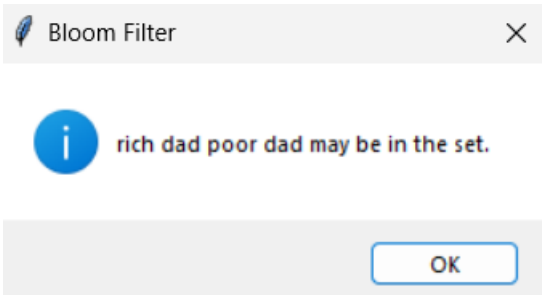




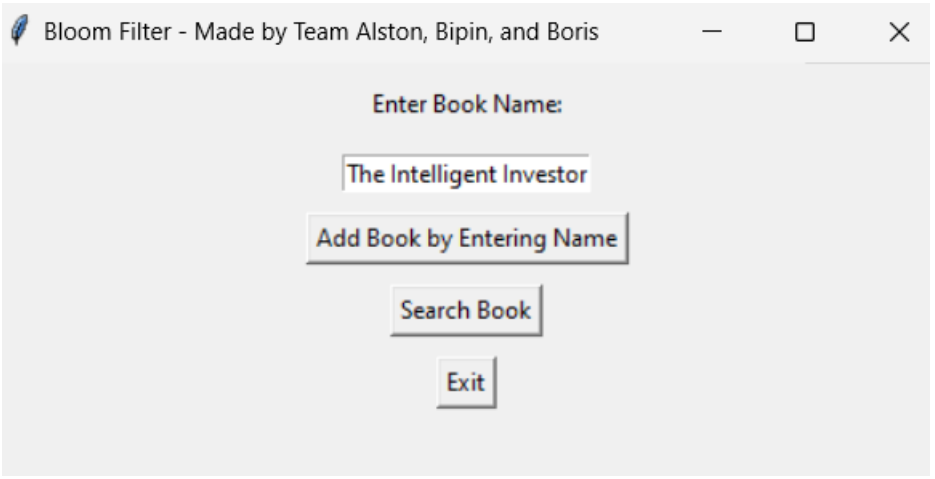


Searching Book:





Searching book not present in library:



Conclusion:	Able to implement Bloom filters for real world problems.
References:	Elearn Notes: https://elearn.dbit.in/course/view.php?id=258 Python Documentation: https://docs.python.org/3/

Don Bosco Institute of Technology
Department of Computer Engineering

Academic year – 2023-2024

Big Data Analytics

Assessment Rubric for Experiment No. : 04

Performance Date :
Submission Date :

Title of Experiment : Implement Bloom Filter using any programming language

Year and Semester : IV^{rh} Year and VIIth Semester

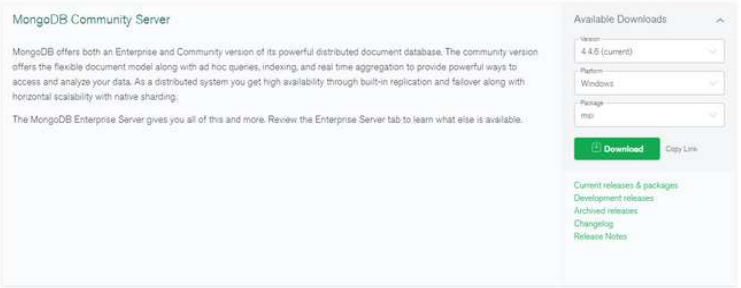
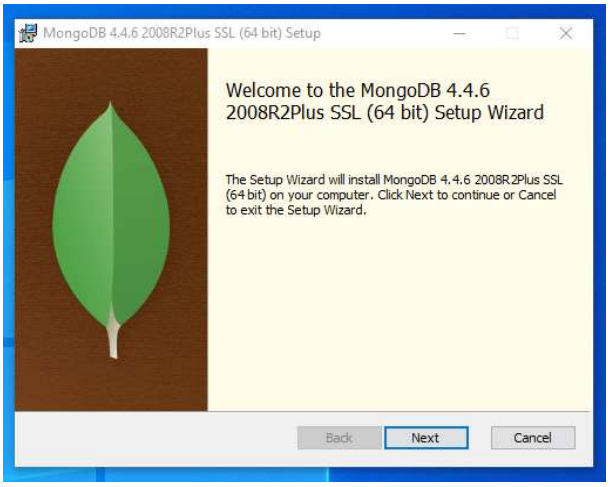
Batch : B

Name of Student: Bipin Dinesh Giri

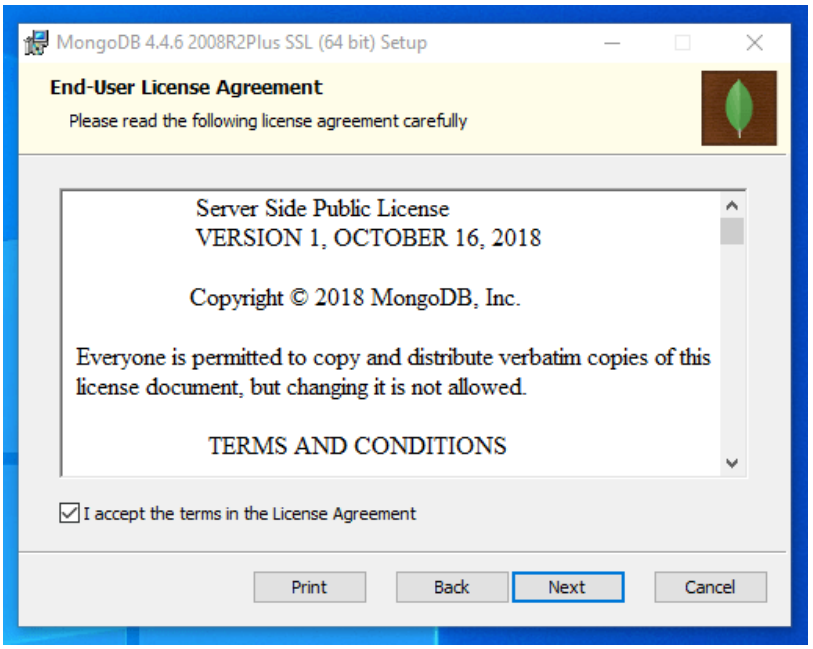
Roll No. : 24

Performance	**Poor**	**Satisfactory**	**Good**	**Excellent**
	1 point	2 points	3 points	4 points
Results and Documentations	**Poor**	**Satisfactory**	**Good**	**Excellent**
	1 point	2 points	3 points	4 points
Viva	**Poor**	**Satisfactory**	**Good**	**Excellent**
	1 point	2 points	3 points	4 points
Timely Submission	**Submission beyond 7 days of the deadline**	**Late submission till 7 days**	**Submission on time**	
	1 points	2 points	3 points	

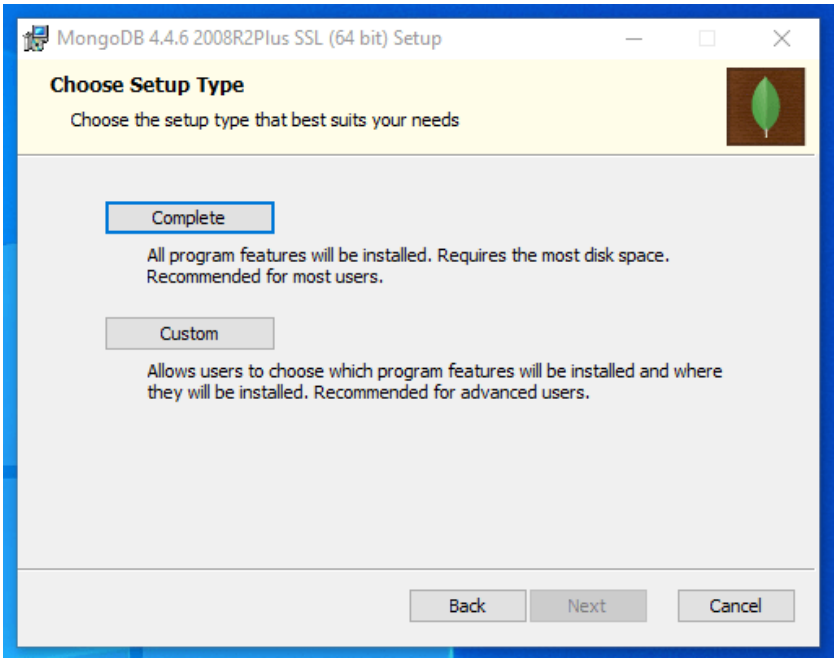
Experiment No: 5

Topic:	To install and configure MongoDB to execute NoSQL commands.
Prerequisite:	Basic knowledge of database, SQL, NoSQL
Mapping With COs:	**CSL702.5**
Objective:	Able to collect, manage, store, query and analyze various forms of Big Data.
Outcomes:	Students will be able to install and manage big data using NoSQL database (Mongodb).
Instructions:	This experiment is a compulsory experiment. All the students are required to perform this experiment individually.
Deliverables:	**Submission:** 1. Installation steps snapshots.  

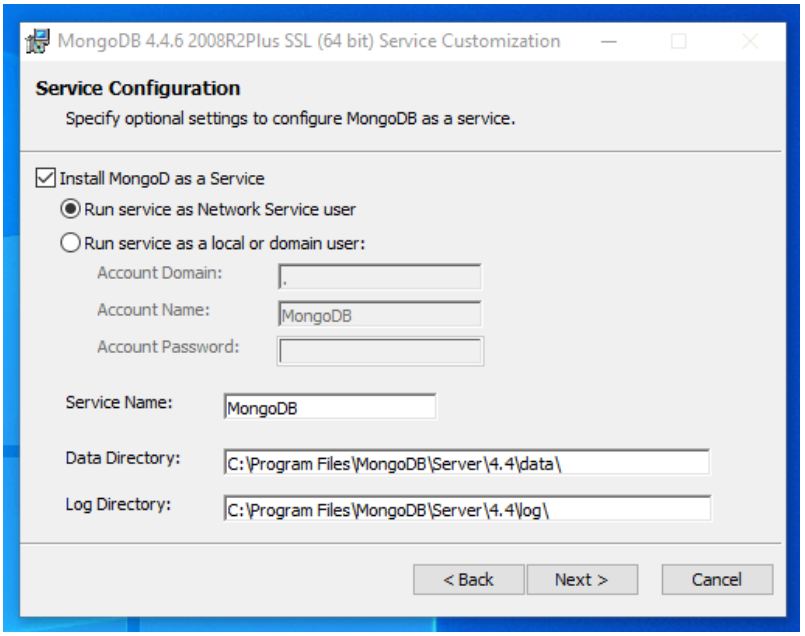
Click **Next** on the initial page to continue.



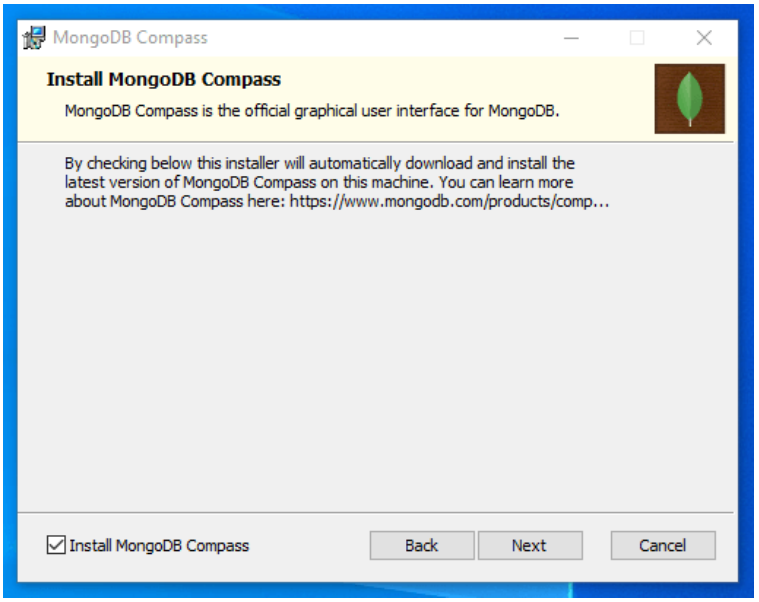
Click **Next** to continue.



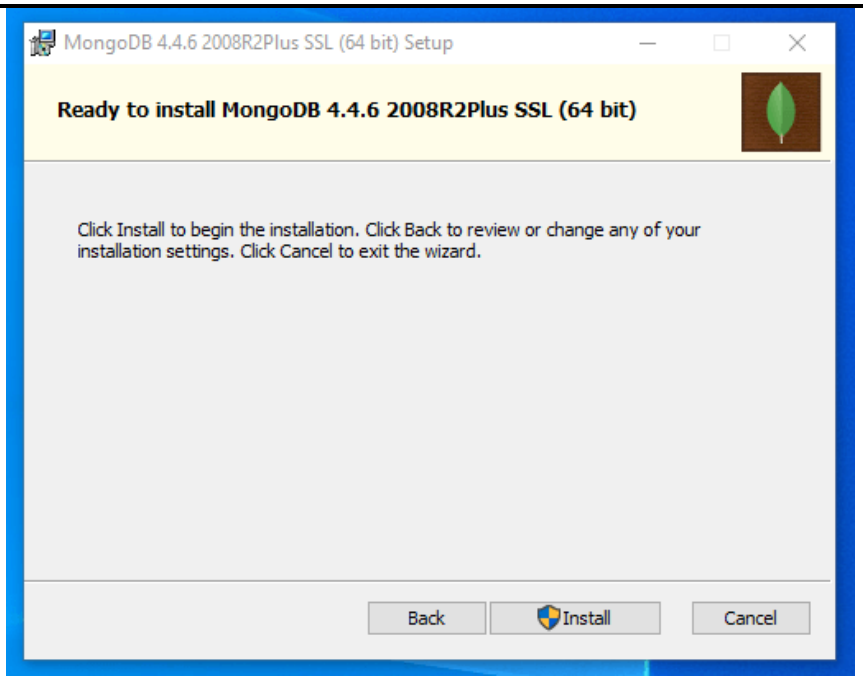
Choose the **Complete** installation to install all of the MongoDB components.



The default values should work well for most scenarios. Click **Next** when you are satisfied with your selections.



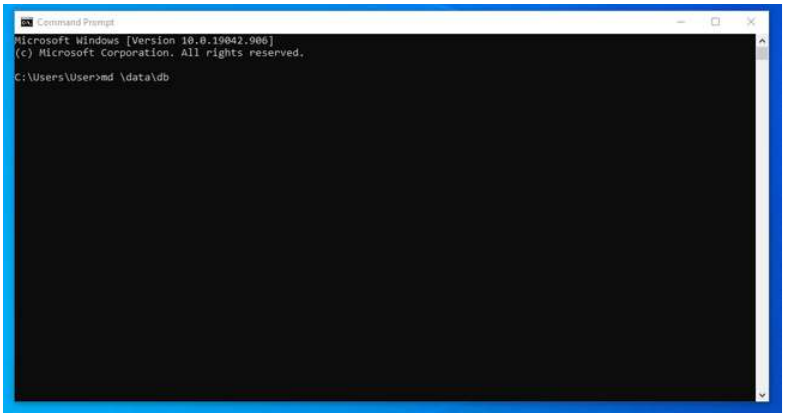
Click **Next** after making your decision.



Click **Install** to begin installing all of the MongoDB components on your computer.

Before you run the server, you need to create the default directory where MongoDB stores its data: `\data\db`. You can create that directory by typing:

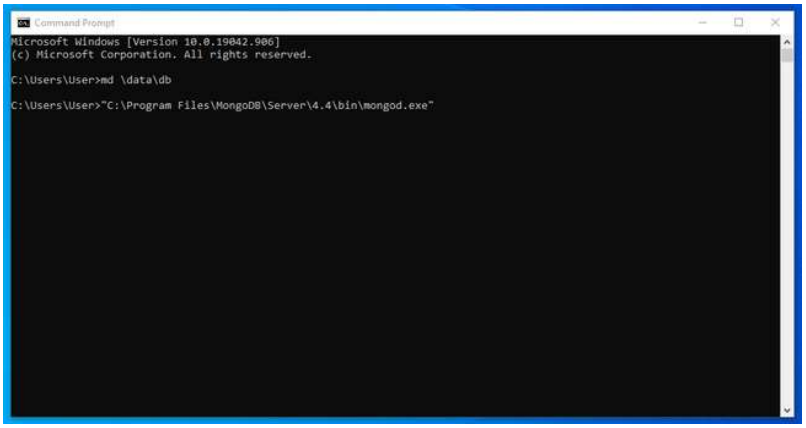
```
md \data\db
```



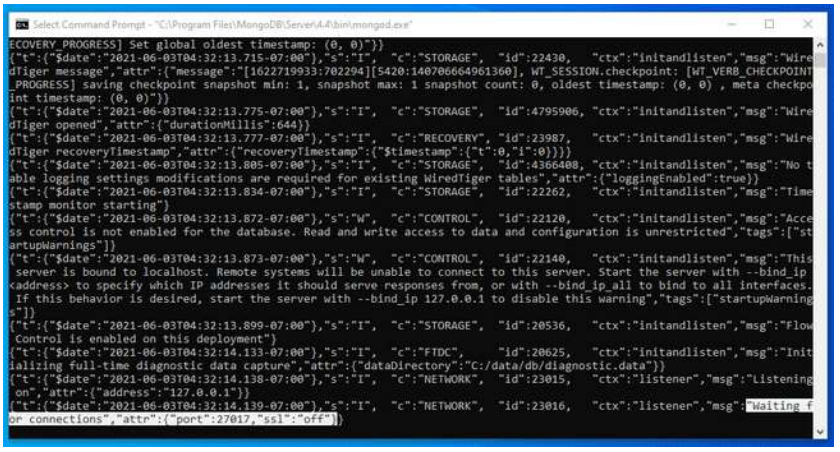
Part of the path contains the MongoDB version number that you installed, so your installation path may be slightly different than the one used below:

```
C:\Program
```

`Files\MongoDB\Server\4.4\bin\mongod.exe`



If everything is functioning correctly, the server will start up and output diagnostic information to the console. To verify that the startup was successful, look for a message that indicates that it is now accepting connections from clients:

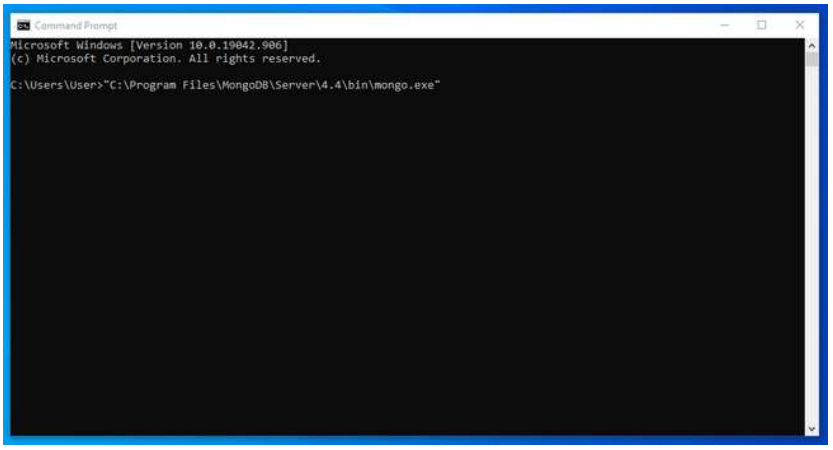


To connect to your running MongoDB server, open another Command Prompt window. Similar to before, we need to type in the absolute path to the executable file.

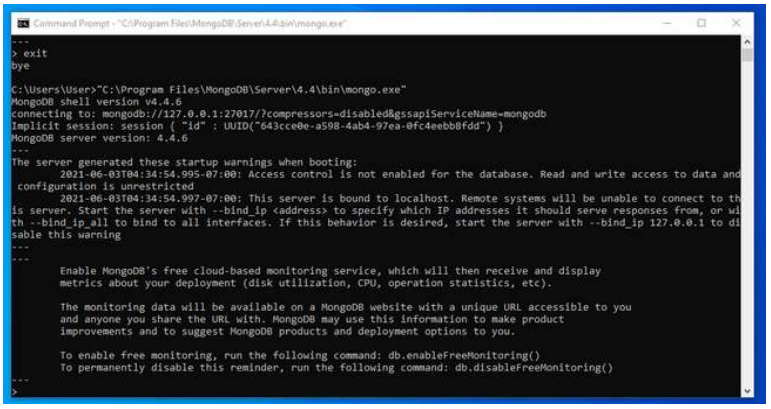
In this case, we are trying to run the `mongo.exe` executable so, taking into account the differences in version numbers, the command should

look something like this:

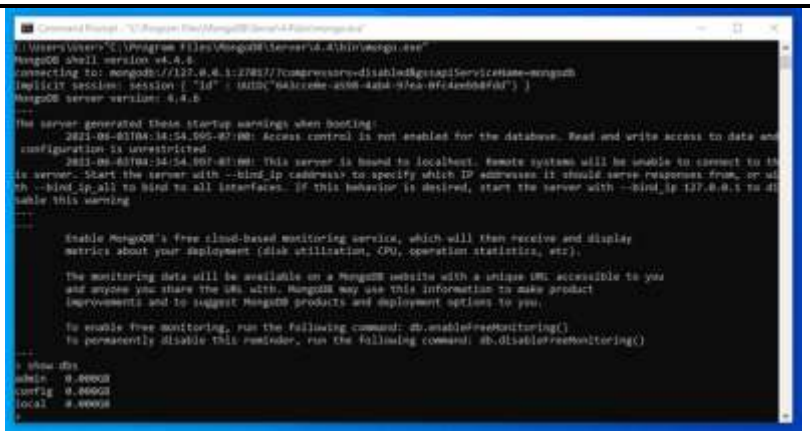
```
C:\Program Files\MongoDB\Server\4.4\bin\mongo.exe
```



Once the shell connects to the server, it will print information about the connection and drop you into a MongoDB prompt:



To verify that the server is responding to commands, run the `show dbs` command:



2. Problem statement (for which domain database is going to maintained)

Introduction:
In the fast-paced world of healthcare, efficient management of patient information and doctor records is paramount for ensuring quality patient care and streamlining hospital operations. To address this need, we are tasked with developing a Hospital Management System using MongoDB as the database system. The system will encompass two main entities: Patients and Doctors. This problem statement outlines the purpose and objectives of the database, as well as the utilization of MongoDB to create a robust Hospital Management System.

Problem Statement:
The primary objective of this project is to design a MongoDB database for a Hospital Management System, which will facilitate efficient storage, retrieval, and manipulation of data related to patients and doctors in a hospital setting. The database should meet the following requirements:

Patient Management:
Store detailed information about patients, including but not limited to their personal details (name, contact information, date of birth), medical history, and admission details.
Enable efficient search and retrieval of patient records for quick access by medical staff, such as nurses and doctors.
Support the addition, modification, and deletion of patient records as needed.

Doctor Management:
Maintain a comprehensive database of doctors working in the hospital, including their contact information, specialization, and availability.
Enable quick access to doctor profiles for appointment scheduling and patient referrals.
Allow the hospital administration to manage doctor records efficiently.

Patient-Doctor Relationship:

Establish a connection between patients and their attending doctors.
Each patient should be associated with a doctor responsible for their treatment.
Enable easy assignment and reassignment of doctors to patients as required.
Store records of appointments and treatments associated with specific patient-doctor relationships.

3. Apply various basic operations and CRUD operations (snapshots of each)

1. Show Databases -

   Command -

   show dbs

```
------

test> show dbs
admin   40.00 KiB
config  12.00 KiB
local   40.00 KiB
test>
```

2. Create Database -

   Command -

   use <database_name>

```
test> use hospital
switched to db hospital
hospital> show dbs
admin   40.00 KiB
config  60.00 KiB
local   40.00 KiB
hospital>
```

3. Create Tables -

   Command -

   db.createCollection("<tablename>")

```
hospital> db.createCollection("doctors")
{ ok: 1 }
hospital> db.createCollection("patients")
```

4. Insert Values in DB

   Command (insertOne)-

   db.<table_name>.insertOne({<key>:"<value>"});

```
hospital> db.Patient.insertOne({
...   ID: 101,
...   Name: "Ella Davis",
...   Place: "Houston",
...   Condition: "Infection"
... });
{
  acknowledged: true,
  insertedId: ObjectId("6527ed561ecdbbd61542461a")
}
```

   Command (insertMany)-

   db.<table_name>.insertMany(
   {<key1>:"<value1>"},
   {<key2>:"<value2>"}
   );

```
hospital> db.Patient.insertMany([
...   {
...     ID: 102,
...     Name: "John Doe",
...     Place: "New York",
...     Condition: "Fever"
...   },
...   {
...     ID: 103,
...     Name: "Jane Smith",
...     Place: "Los Angeles",
...     Condition: "Fracture"
...   },
...   {
...     ID: 104,
...     Name: "Alice Johnson",
...     Place: "Chicago",
...     Condition: "Allergy"
...   }
hospital>
{
  acknowledged: true,
  insertedIds: {
    '0': ObjectId("6527ed791ecdbbd61542461b"),
    '1': ObjectId("6527ed791ecdbbd61542461c"),
    '2': ObjectId("6527ed791ecdbbd61542461d")
  }
```

5. Display -

   Command - db.<table_name>.find()

```
hospital> db.Patient.find()
[
  {
    _id: ObjectId("6527ed561ecdbbd61542461a"),
    ID: 101,
    Name: 'Ella Davis',
    Place: 'Houston',
    Condition: 'Infection'
  },
  {
    _id: ObjectId("6527ed791ecdbbd61542461b"),
    ID: 102,
    Name: 'John Doe',
    Place: 'New York',
    Condition: 'Fever'
  },
  {
    _id: ObjectId("6527ed791ecdbbd61542461c"),
    ID: 103,
    Name: 'Jane Smith',
    Place: 'Los Angeles',
    Condition: 'Fracture'
  },
  {
    _id: ObjectId("6527ed791ecdbbd61542461d"),
    ID: 104,
    Name: 'Alice Johnson',
    Place: 'Chicago',
    Condition: 'Allergy'
  }
]
```

6. Update -

   Command -

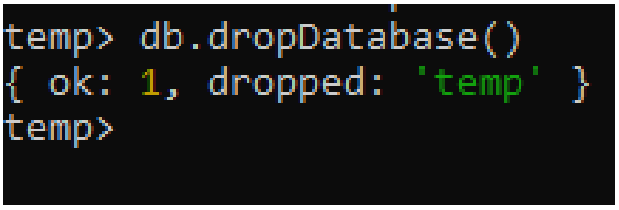
   db.<table_name>.updateOne(&set:{<key>:<new_value>})

```
hospital> db.Patient.updateOne(
...    { ID: 101 },
...    { $set: { Name: "Michael Jackson" } }
... );
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
hospital> db.Patient.find()
[
  {
    _id: ObjectId("6527ed561ecdbbd61542461a"),
    ID: 101,
    Name: 'Michael Jackson',
    Place: 'Houston',
    Condition: 'Infection'
  },
  {
    _id: ObjectId("6527ed791ecdbbd61542461b"),
    ID: 102,
    Name: 'John Doe',
    Place: 'New York',
    Condition: 'Fever'
  },
  {
    _id: ObjectId("6527ed791ecdbbd61542461c"),
    ID: 103,
    Name: 'Jane Smith',
    Place: 'Los Angeles',
    Condition: 'Fracture'
  },
  {
    _id: ObjectId("6527ed791ecdbbd61542461d"),
    ID: 104,
    Name: 'Alice Johnson',
    Place: 'Chicago',
    Condition: 'Allergy'
  }
]
```

   Delete Table -

   db.<table_name>.drop()

```
hospital> db.Doctor.drop()
true
```

	Drop Database - 
Conclusion:	Thus students will be able to successfully installed mongodb and applied CRUD operations to manage the big data.
References:	https://www.mongodb.com/ https://docs.mongodb.com/ https://university.mongodb.com/

Don Bosco Institute of Technology
Department of Computer Engineering

Academic year – 2023-2024

Big Data Analytics
Assessment Rubric for Experiment No. : 05

Performance Date :
Submission Date :

Title of Experiment : To install and configure MongoDB to execute NoSQL commands.

Year and Semester : IV^{rh} Year and VIIth Semester

Batch : A

Name of Student : Alston Fernandes

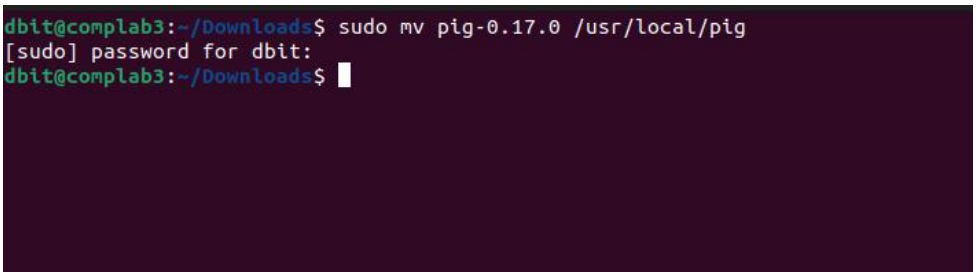
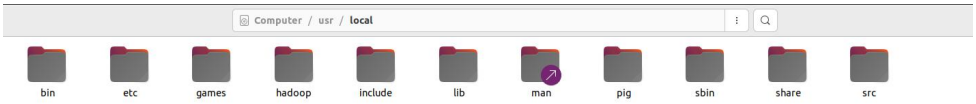
Roll No. : 19

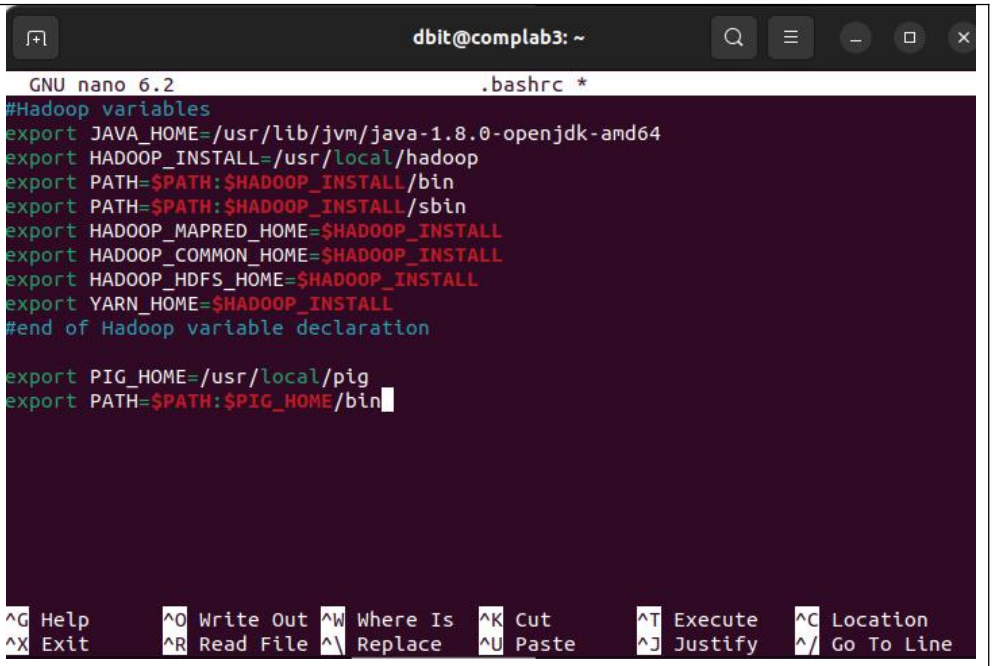
Performance	**Poor**	**Satisfactory**	**Good**	**Excellent**
	1 point	2 points	3 points	4 points
Results and Documentations	**Poor**	**Satisfactory**	**Good**	**Excellent**
	1 point	2 points	3 points	4 points
Viva	**Poor**	**Satisfactory**	**Good**	**Excellent**
	1 point	2 points	3 points	4 points
Timely Submission	**Submission beyond 7 days of the deadline**	**Late submission till 7 days**	**Submission on time**	
	1 points	2 points	3 points	

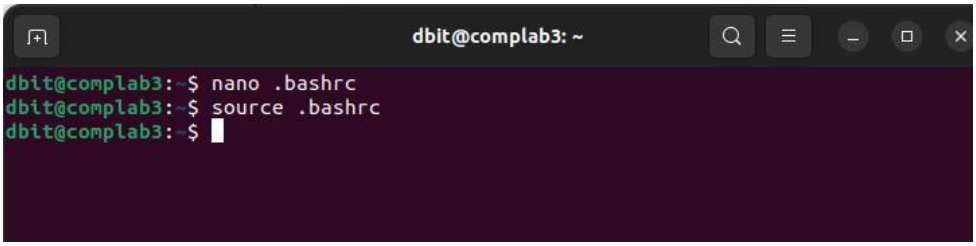
Faculty Incharge : Ms. Sana Shaikh

Experiment No: 6

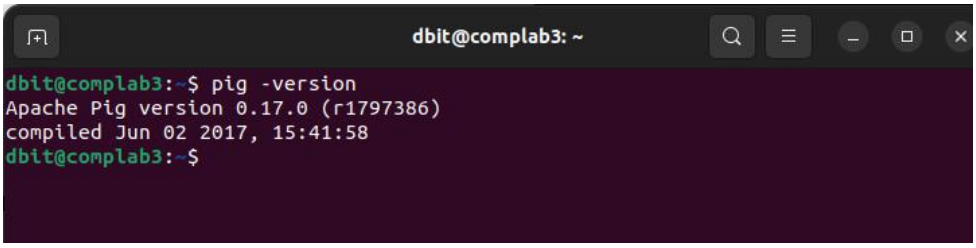
Name: Alston Fernandes **Roll No:** 19 **Batch:** A

Topic:	Install, configure and execute Apache PIG Latin commands
Mapping With COs:	CSL702.3
Objective:	To write queries in PIG Latin language to filter data.
Outcomes:	Students will be able to install and manage big data using Apache PIG.
Instructions:	This experiment is a compulsory experiment. All the students are required to perform this experiment individually.
Deliverables:	**Installing PIG**   **Updating bashrc file**





Checking PIG is installed correctly



```
dbit@complab3: ~
dbit@complab3:~$ pig
23/09/23 18:33:49 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
23/09/23 18:33:49 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
23/09/23 18:33:49 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2023-09-23 18:33:49,635 [main] INFO  org.apache.pig.Main - Apache Pig version 0.
17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2023-09-23 18:33:49,635 [main] INFO  org.apache.pig.Main - Logging error message
s to: /home/dbit/pig_1695474229633.log
2023-09-23 18:33:49,767 [main] INFO  org.apache.pig.impl.util.Utils - Default bo
otup file /home/dbit/.pigbootup not found
2023-09-23 18:33:51,037 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2023-09-23 18:33:51,037 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: hdfs://localhost:9000
2023-09-23 18:33:52,622 [main] INFO  org.apache.pig.PigServer - Pig Script ID fo
r the session: PIG-default-3d85f299-a1da-45bb-9f6a-5fa7bc5fb011
2023-09-23 18:33:52,622 [main] WARN  org.apache.pig.PigServer - ATS is disabled
since yarn.timeline-service.enabled set to false
grunt>
```

Dataset 1

```
dbit@complab3:~$ cat pig.txt
1991, 33, Mumbai
1991, 38, Pune
1991, 38, Pune
1991, 36, Delhi
1990, 55, Pune
1990, 55, Delhi
1996, 65, Mumbai
1996, 36, Pune
2020, 66, Mumbai
2020, 52, Pune
2025, 55, Mumbai
2025, 62, Delhi
2025, 62, Pune
2025, 62, Pune
dbit@complab3:~$
```

LOAD keyword with column name and datatype

```
grunt> my_bag = LOAD '/home/dbit/pig.txt' as (year:int, temp:int, city:chararray);
2023-09-23 18:46:24,241 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> describe my_bag;
my_bag: {year: int,temp: int,city: chararray}
grunt>
```

Dump Keyword

```
2023-09-24 11:52:22,149 [main] INFO  org.apache.hadoop.conf.Configurat
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per
grunt> dump my_bag;
```

```
hemaTupleBackend has already been initialized
2023-09-24 11:53:02,166 [main] INFO  org.apache.hadoop.mapreduce.lib.in
nputFormat - Total input paths to process : 1
2023-09-24 11:53:02,166 [main] INFO  org.apache.pig.backend.hadoop.exec
ne.util.MapRedUtil - Total input paths to process : 1
(1991, 33, Mumbai)
(1991, 38, Pune)
(1991, 38, Pune)
(1991, 36, Delhi)
(1990, 55, Pune)
(1990, 55, Delhi)
(1996, 65, Mumbai)
(1996, 36, Pune)
(2020, 66, Mumbai)
(2020, 52, Pune)
(2025, 55, Mumbai)
(2025, 62, Delhi)
(2025, 62, Pune)
(2025, 62, Pune)
grunt>
```

LOAD keyword with column name

```
grunt> my_bag1 = LOAD '/home/dbit/pig.txt' as (year, temp, city);
2023-09-23 19:23:54,441 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> describe my_bag1;
my_bag1: {year: bytearray,temp: bytearray,city: bytearray}
grunt>
```

LOAD keyword without parameter

```
grunt> my_bag2 = LOAD '/home/dbit/pig.txt';
2023-09-23 19:24:36,894 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> describe my_bag2;
Schema for my_bag2 unknown.
grunt>
```

Dataset 2

```
dbit@complab3:~$ cat data.txt
Sam, Mumbai, 66
Jim, Pune, 77
Tom, Pune, 55
Herry, Mumbai, 65
Sony, Delhi, 51
Sia, Pune, 65
Sia, Pune, 65
Sia, Pune, 65
Tina, Mumbai, 53
Jia, Delhi, 76
Sara, Delhi, 65
dbit@complab3:~$
```

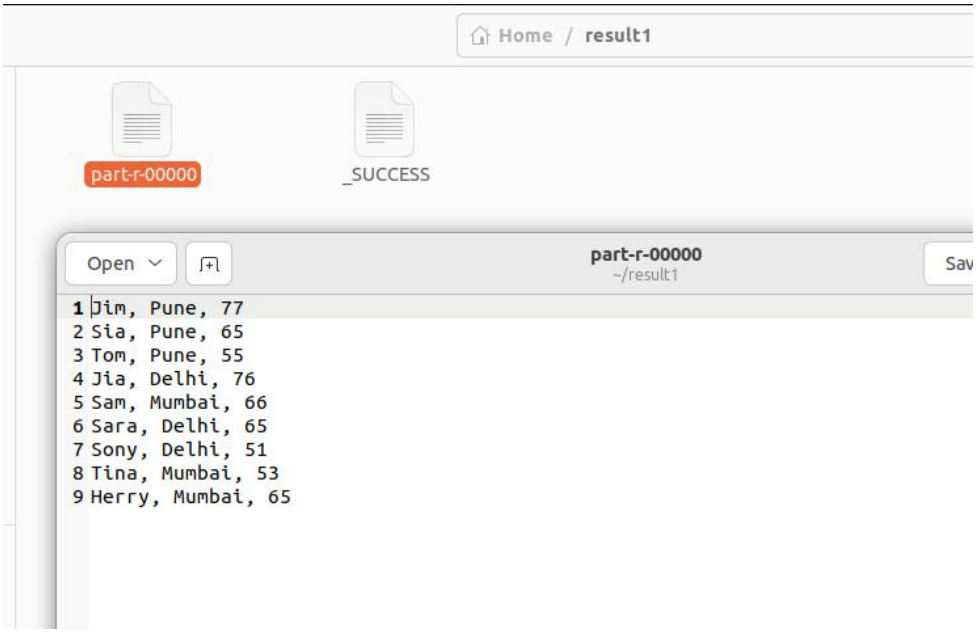
Q1. Remove duplicate tuples and put in a bag say “result1” and show its content.

```
grunt> Q1 = LOAD '/home/dbit/data.txt';
2023-09-23 19:28:03,678 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> result1 = DISTINCT Q1;
grunt> dump result1;
2023-09-23 19:29:04,365 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig fe
d in the script: DISTINCT
2023-09-23 19:29:04,389 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-23 19:29:04,389 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTup
has already been initialized
2023-09-23 19:29:04,390 [main] INFO  org.apache.pig.newplan.logical.optimizer.LogicalPl
```

```
2023-09-24 11:55:04,645 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2023-09-24 11:55:04,645 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(Jim, Pune, 77)
(Sia, Pune, 65)
(Tom, Pune, 55)
(Jia, Delhi, 76)
(Sam, Mumbai, 66)
(Sara, Delhi, 65)
(Sony, Delhi, 51)
(Tina, Mumbai, 53)
(Herry, Mumbai, 65)
grunt>
```

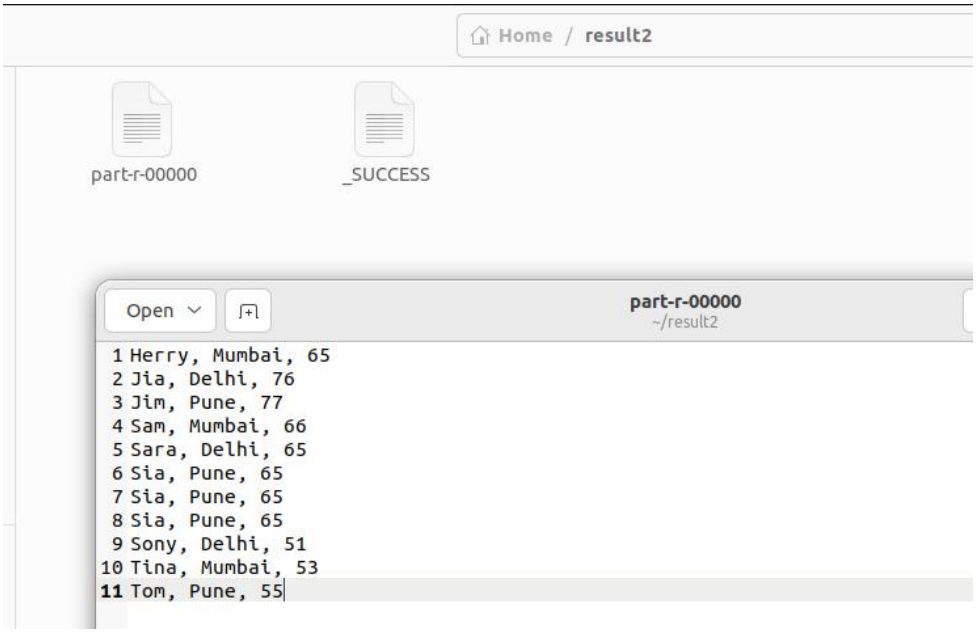
```
grunt> store result1 into '/home/dbit/result1' using PigStorage(';');
2023-09-24 11:56:39,360 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-24 11:56:39,364 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - mapred.textoutputformat.separator is deprecated. Instead, use mapreduce.
output.textoutputformat.separator
2023-09-24 11:56:39,400 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: DISTINCT
2023-09-24 11:56:39,416 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-24 11:56:39,416 [main] WARN  org.apache.pig.data.SchemaTupleBackend - Sc
hemaTupleBackend has already been initialized
2023-09-24 11:56:39,416 [main] INFO  org.apache.pig.newplan.logical.optimizer.Lo
gicalPlanOptimizer - {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, ConstantCalc
ulator, GroupByConstParallelSetter, LimitOptimizer, LoadTypeCastInserter, MergeF
```

```
Job DAG:
job_local1902519006_0003


2023-09-23 19:31:06,568 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize
 JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-09-23 19:31:06,568 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize
 JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-09-23 19:31:06,569 [main] INFO  org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize
 JVM Metrics with processName=JobTracker, sessionId= - already initialized
2023-09-23 19:31:06,570 [main] INFO  org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.MapReduceLauncher - Success!
grunt>
```

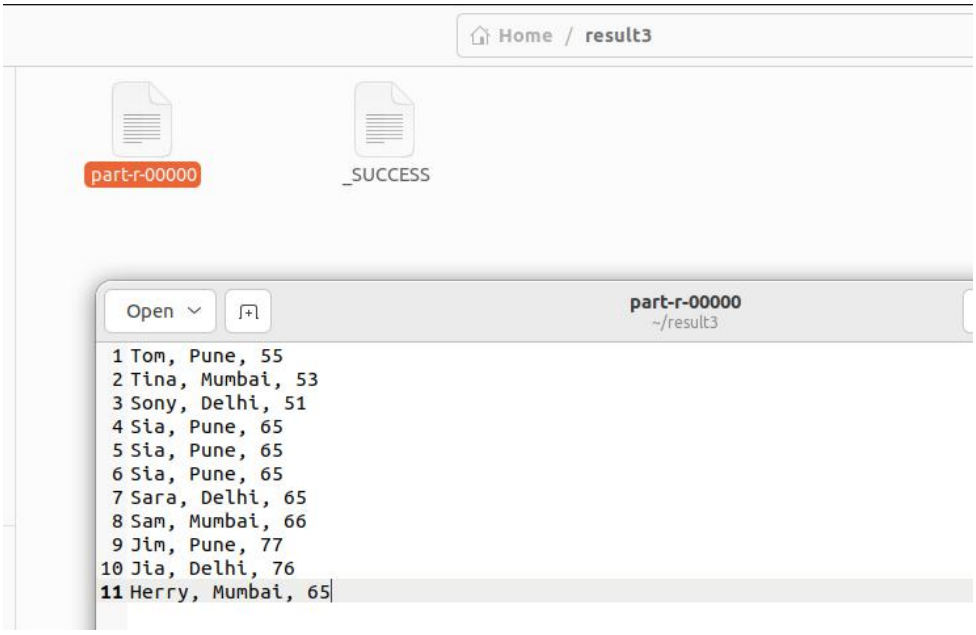


Q2. Sort the data in ascending and descending both and put the sorted data in the bag “result2” and “result3” respectively.

```
grunt> Q2 = LOAD '/home/dbit/data.txt' as (Name, Location, marks);
2023-09-23 19:37:02,550 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.byte
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> describe Q2;
Q2: {Name: bytearray,Location: bytearray,marks: bytearray}
grunt> result2 = ORDER Q2 by Name;
grunt> result3 = ORDER Q2 by Name DESC;
grunt> store result2 into '/home/dbit/result2' using PigStorage(';');
2023-09-23 19:38:12,021 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.byte
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-23 19:38:12,062 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features
d in the script: ORDER_BY
```




```
grunt> store result3 into '/home/dbit/result3' using PigStorage(';');
2023-09-23 19:39:21,781 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.byte
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-23 19:39:21,815 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.byte
```

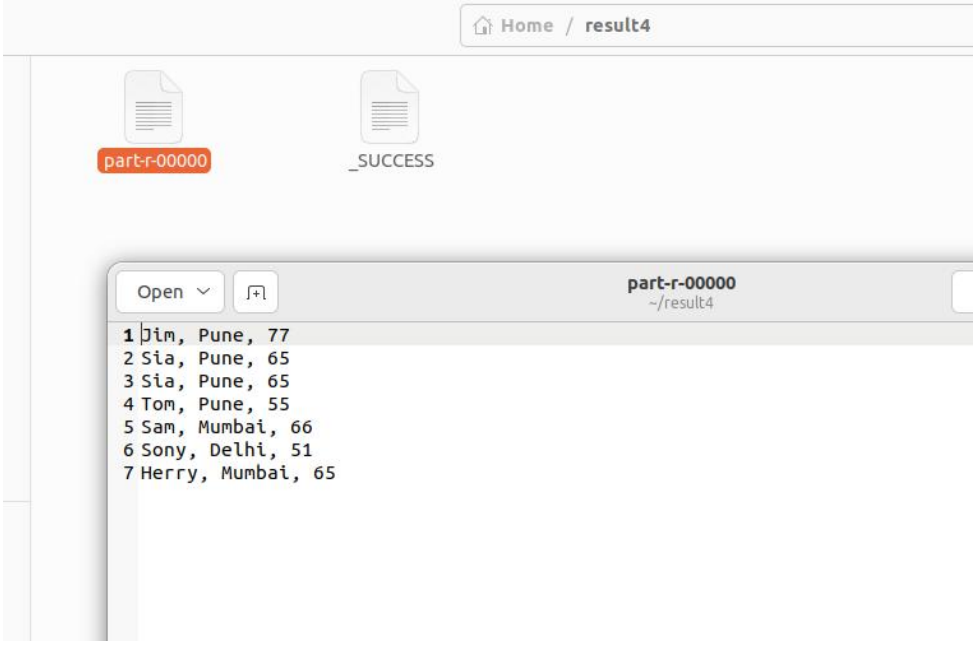


Q3. Put the first seven tuples in the bag “result4” and show its content.

```
grunt> Q3 = LOAD '/home/dbit/data.txt';
2023-09-23 21:02:03,992 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-23 21:02:04,249 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> result4 = LIMIT Q3 7;
grunt> dump result4;
```

```
hemaTupleBackend has already been initialized
2023-09-24 12:07:21,069 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileI
nputFormat - Total input paths to process : 1
2023-09-24 12:07:21,069 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(Sam, Mumbai, 66)
(Jim, Pune, 77)
(Tom, Pune, 55)
(Herry, Mumbai, 65)
(Sony, Delhi, 51)
(Sia, Pune, 65)
(Sia, Pune, 65)
grunt>
```

```
grunt> store result4 into '/home/dbit/result4' using PigStorage(';');
2023-09-23 21:03:32,695 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-23 21:03:33,075 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features use
d in the script: LIMIT
```



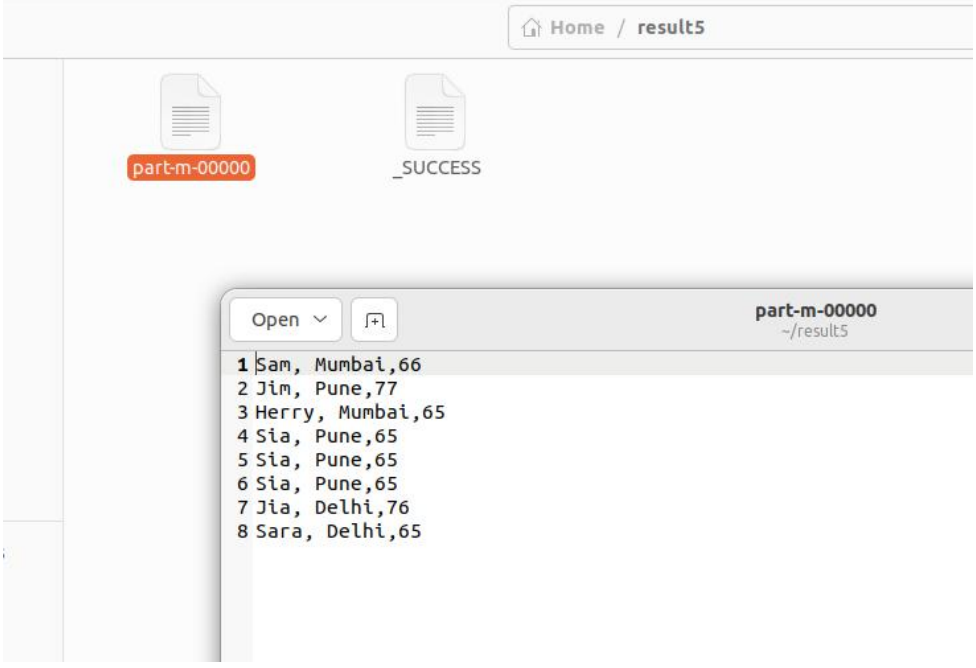
Q4. Show all the details of the students who scored more than 60 marks in “result5”.

```
grunt> Q4 = LOAD '/home/dbit/data.txt' using PigStorage(',') as (Name:chararray,
 Location:chararray, Marks:int);
2023-09-24 12:21:38,527 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> dump Q4;
2023-09-24 12:21:40,086 [main] INFO  org.apache.pig.tools.pigstats.ScriptState -
 Pig features used in the script: UNKNOWN
2023-09-24 12:21:40,097 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
```

```
2023-09-24 12:21:40,412 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(Sam, Mumbai,66)
(Jim, Pune,77)
(Tom, Pune,55)
(Herry, Mumbai,65)
(Sony, Delhi,51)
(Sia, Pune,65)
(Sia, Pune,65)
(Sia, Pune,65)
(Tina, Mumbai,53)
(Jia, Delhi,76)
(Sara, Delhi,65)
grunt> result5 = FILTER Q4 by Marks>60;
grunt> dump result5;
```

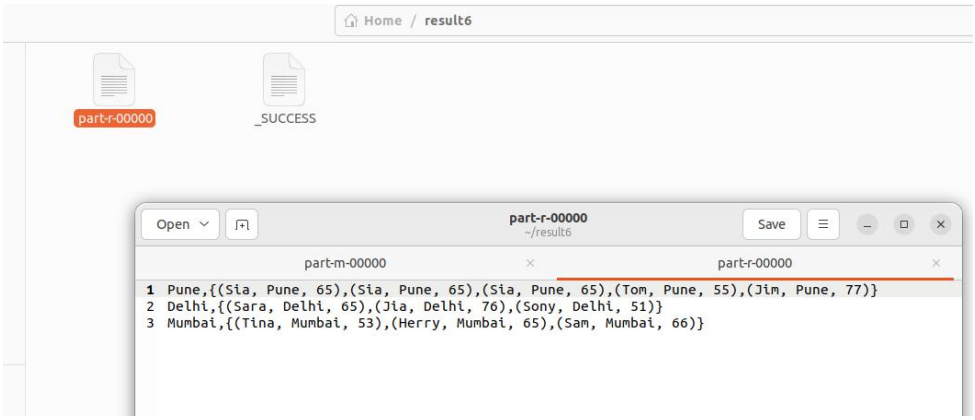
```
nputFormat - Total input paths to process : 1
2023-09-24 12:23:31,876 [main] INFO  org.apache.pig.backend.hadoop.executionengi
ne.util.MapRedUtil - Total input paths to process : 1
(Sam, Mumbai,66)
(Jim, Pune,77)
(Herry, Mumbai,65)
(Sia, Pune,65)
(Sia, Pune,65)
(Sia, Pune,65)
(Jia, Delhi,76)
(Sara, Delhi,65)
grunt>
```

```
grunt> store result5 into '/home/dbit/result5' using PigStorage(',');
2023-09-24 12:24:51,066 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-24 12:24:51,081 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
```



Q5. Group the students according to their city and store it in the bag saying “result6”.

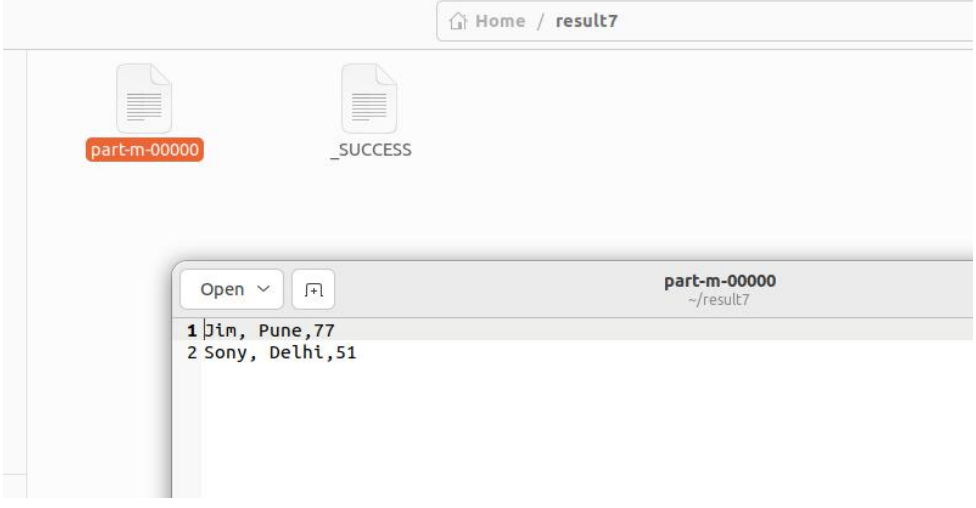
```
grunt> Q5 = LOAD '/home/dbit/data.txt' using PigStorage(',') as (Name, Location:
chararray, Marks);
2023-09-24 12:26:58,773 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-24 12:26:58,788 [main] INFO  org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> result6 = GROUP Q5 by Location;
grunt> store result6 into '/home/dbit/result6' using PigStorage(',');
```



Q6. Show which student scored maximum and which student scored minimum marks in “result7”

```
grunt> Q6 = LOAD '/home/dbit/data.txt' USING PigStorage(',')
>> AS (Name:chararray, Location:chararray, Marks:int);
2023-09-24 13:38:01,415 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
 io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> min_score = FOREACH (GROUP Q6 ALL) GENERATE MIN(Q6.Marks) AS (Marks:int);
grunt> max_score = FOREACH (GROUP Q6 ALL) GENERATE MAX(Q6.Marks) AS (Marks:int);
grunt> result7 = FILTER Q6 by (Marks==min_score.Marks OR Marks==max_score.Marks);
grunt> dump result7;
2023-09-24 13:38:41,370 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig fe
atures used in the script: GROUP_BY,FILTER
2023-09-24 13:38:41,381 [main] INFO  org.apache.hadoop.conf.Configuration.deprecation -
 io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2023-09-24 13:38:41,382 [main] INFO  org.apache.pig.data.SchemaTupleBackend - Key [pig
```

```
2023-09-24 13:38:42,118 [main] WARN  org.apache.pig.data.SchemaTupleBackend - SchemaTup
leBackend has already been initialized
2023-09-24 13:38:42,124 [main] INFO  org.apache.hadoop.mapreduce.lib.input.FileInputFor
mat - Total input paths to process : 1
2023-09-24 13:38:42,124 [main] INFO  org.apache.pig.backend.hadoop.executionengine.util
.MapRedUtil - Total input paths to process : 1
(Jim, Pune,77)
(Sony, Delhi,51)
grunt>
grunt>
```



Conclusion:	Apache Pig offers a high-level scripting language that abstracts complex Hadoop MapReduce operations. This simplicity makes it an excellent tool for teaching beginners or students who are new to big data processing and distributed computing. Pig seamlessly integrates with other components of the Hadoop ecosystem, such as HDFS (Hadoop Distributed File System) and Hive. This integration exposes students to a broader ecosystem of tools and technologies used for big data processing and analytics.
References:	https://pig.apache.org/docs/latest/basic.html#schemas https://www.tutorialspoint.com/apache_pig/pig_latin_basics.htm https://www.tutorialspoint.com/apache_pig/apache_pig_min.htm https://data-flair.training/blogs/apache-pig-built-in-functions/

Don Bosco Institute of Technology
Department of Computer Engineering

Academic year – 2022-2023

Big Data Analytics

Assessment Rubric for Experiment No.: 06

Performance Date :
Submission Date :

Title of Experiment :

Year and Semester : IV^{rh} Year and VIIth Semester
Batch :
Name of Student :
Roll No. :

Performance	**Poor**	**Satisfactory**	**Good**	**Excellent**	
	1 point	2 points	3 points	4 points	
Results and Documentations	**Poor**	**Satisfactory**	**Good**	**Excellent**	
	1 point	2 points	3 points	4 points	
Viva	**Poor**	**Satisfactory**	**Good**	**Excellent**	
	1 point	2 points	3 points	4 points	
Timely Submission	**Submission beyond 7 days of the deadline**	**Late submission till 7 days**	**Submission on time**		
	1 points	2 points	3 points		

Faculty Incharge : Ms. Sana Shaikh

Name: Alston Fernandes **Roll no.: 19** **Batch: A**

Experiment No: 7 & 8

Topic:	Data Analysis & Data Visualization using Hive/PIG/R/Tableau.
Prerequisite:	Basic knowledge of Data Analysis & data visualization tools like Hive/PIG/R/Tableau etc.
Mapping With COs:	**CSL702.6**
Objective:	Able to be used for data analysis, statistical modeling, and machine learning tasks.
Outcomes:	Analyze the datasets using Data Analysis tools & generate reports using data visualization tools capabilities.
Instructions:	This experiment is a compulsory experiment. All the students are required to perform this experiment individually.
Deliverables:	**Submission:** 1. Installation steps snapshots for Windows 10 systems Step 1: Search **Download Tableau** in Internet Browser, you will be shown with the installation page on the first link or use the following link - https://www.tableau.com/products/desktop/download  You will be redirected to the following page Step 2: Click on Start your free trial button to install in your Windows System.

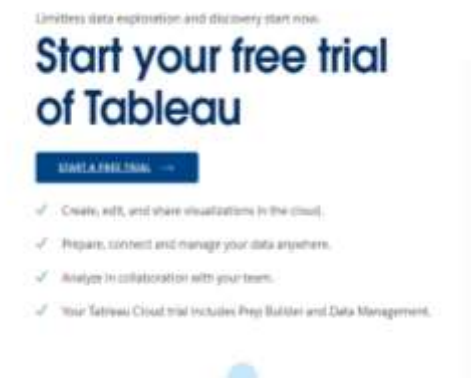
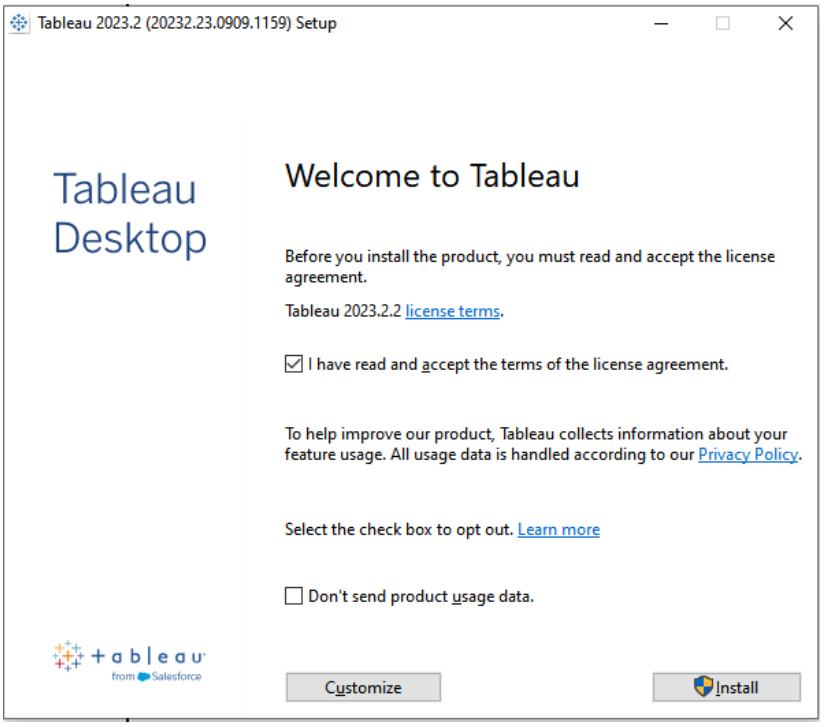


Tableau software will get download into your local system.

Step 3: Installing the software in local machine -



Step 4 - Tableau will be installed in the system and will be redirected with the page below -

2. Detail about Datasets

Dataset used for visualization

1. Customers Database -

customer_code	custmer_name	customer_type
Cus001	Surge Stores	Brick & Mortar
Cus002	Nomad Stores	Brick & Mortar
Cus003	Excel Stores	Brick & Mortar
Cus004	Surface Stores	Brick & Mortar
Cus005	Premium Stores	Brick & Mortar
Cus006	Electricalsara Stores	Brick & Mortar
Cus007	Info Stores	Brick & Mortar
Cus008	Acclaimed Stores	Brick & Mortar
Cus009	Electricalsquipo Stores	Brick & Mortar
Cus010	Atlas Stores	Brick & Mortar
Cus011	Flawless Stores	Brick & Mortar
Cus012	Integration Stores	Brick & Mortar
Cus013	Unity Stores	Brick & Mortar
Cus014	Forward Stores	Brick & Mortar
Cus015	Electricalsbea Stores	Brick & Mortar
Cus016	Logic Stores	Brick & Mortar
Cus017	Epic Stores	Brick & Mortar
Cus018	Electricalslance Stores	Brick & Mortar

2. Market's Database -

markets_code	markets_name	zone
Mark001	Chennai	South
Mark002	Mumbai	Central
Mark003	Ahmedabad	North
Mark004	Delhi NCR	North
Mark005	Kanpur	North
Mark006	Bengaluru	South
Mark007	Bhopal	Central
Mark008	Lucknow	North
Mark009	Patna	North
Mark010	Kochi	South
Mark011	Nagpur	Central
Mark012	Surat	North
Mark013	Bhopal	Central
Mark014	Hyderabad	South
Mark015	Bhubaneshwar	South
Mark097	New York	
Mark999	Paris	

3. Product's Database -

product_code	product_type
Prod001	Own Brand
Prod002	Own Brand
Prod003	Own Brand
Prod004	Own Brand
Prod005	Own Brand
Prod006	Own Brand
Prod007	Own Brand
Prod008	Own Brand
Prod009	Own Brand
Prod010	Own Brand
Prod011	Own Brand
Prod012	Own Brand
Prod013	Own Brand
Prod014	Own Brand
Prod015	Own Brand
Prod016	Own Brand
Prod017	Own Brand
Prod018	Own Brand

4. Transaction's Database -

product_code	customer_code	market_code	order_date	sales_qty	sales_amount	currency
Prod001	Cus001	Mark001	2017-10-10	100	41241	INR
Prod001	Cus002	Mark002	2018-05-08	3	-1	INR
Prod002	Cus003	Mark003	2018-04-06	1	875	INR
Prod002	Cus003	Mark003	2018-04-11	1	583	INR
Prod002	Cus004	Mark003	2018-06-18	6	7176	INR
Prod003	Cus005	Mark004	2017-11-20	59	500	USD
Prod003	Cus005	Mark004	2017-11-22	36	250	USD
Prod003	Cus005	Mark004	2017-11-23	39	21412	INR
Prod003	Cus005	Mark004	2017-11-27	35	19213	INR
Prod003	Cus005	Mark004	2017-11-28	310	170185	INR
Prod003	Cus005	Mark004	2017-11-29	184	101194	INR
Prod003	Cus005	Mark004	2017-11-30	35	19213	INR
Prod004	Cus005	Mark004	2017-11-29	17	9426	INR
Prod004	Cus005	Mark004	2017-12-19	1	218	INR
Prod005	Cus005	Mark004	2018-08-07	5	3093	INR
Prod003	Cus006	Mark004	2017-12-04	58	30306	INR
Prod005	Cus006	Mark004	2018-06-29	38	52319	INR
Prod005	Cus006	Mark004	2018-07-02	93	126296	INR

3. Data Analysis

The data required for this analysis will be sourced from the following tables:

Transactions: Contains detailed information about each sales transaction, including product IDs, customer IDs, market IDs, revenue, and quantity sold.

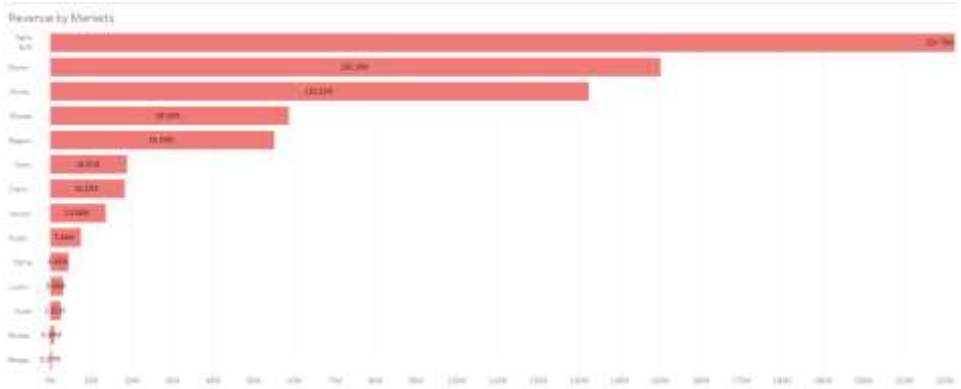
Customers: Contains customer information such as names, addresses, and contact details.

Date: Contains date-related information, allowing for date-based analysis.

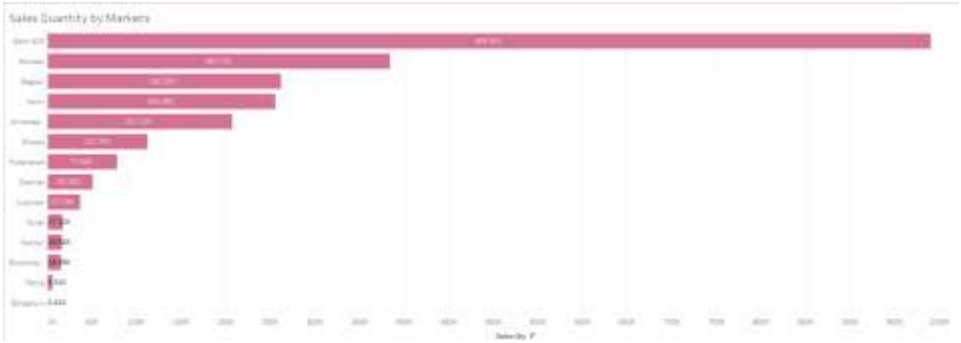
Markets: Provides details about the markets in which Company XYZ operates.

	Products: Contains information about the products Company XYZ sells, including names, descriptions, and pricing. Dashboard will visualize the following requirements - 1. Total Revenue: Calculate the total revenue generated by Company XYZ over the entire dataset period. 2. Total Sales Quantity: Determine the total quantity of products sold by Company XYZ. 3. Revenue by Markets: Display the revenue generated by each market within Company XYZ's operations. 4. Sales Quantity by Markets: Provide a breakdown of the quantity of products sold in each market. 5. Top 5 Customers: Identify and present the top 5 customers who have contributed the most to the company's revenue. 6. Top 5 Products Sold: Identify and present the top 5 products that have been sold the most in terms of quantity. 7. Year-on-Year Sales Insights: ● Calculate and visualize year-on-year total revenue to track revenue growth or decline. ● Analyze year-on-year sales quantity to understand product demand trends over time. ● Provide an option to filter data by specific product IDs and months for more detailed insights. 8. Customer Analysis: Allow users to select specific customers and view their purchase history and contribution to revenue over the years.
	4. Data Visualization **Dashboard -**

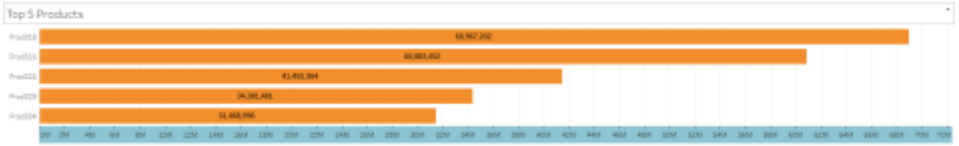
Revenue by Markets -



Sales Quantity by Markets –



Top 5 Customers –

	Top 5 Products Sold –  **Revenue Per Year -** 
Conclusion:	Students will be able to successfully perform Data Analysis & Data Visualization using Hive/PIG/R/Tableau.
References:	Put the reference of resources used to perform this experiment.

Don Bosco Institute of Technology
Department of Computer Engineering

Academic year – 2023-2024

Big Data Analytics

Assessment Rubric for Experiment No. : 07 & 08

Performance Date : 11/10/23
Submission Date : 11/10/23

Title of Experiment : Data Analysis & Data Visualization using Hive/PIG/R/Tableau

Year and Semester : IV^{rh} Year and VIIth Semester
Batch : A
Name of Student : Alston Fernandes
Roll No. : 19

Performance	**Poor**	**Satisfactory**	**Good**	**Excellent**
	1 point	2 points	3 points	4 points
Results and Documentations	**Poor**	**Satisfactory**	**Good**	**Excellent**
	1 point	2 points	3 points	4 points
Viva	**Poor**	**Satisfactory**	**Good**	**Excellent**
	1 point	2 points	3 points	4 points
Timely Submission	**Submission beyond 7 days of the deadline**	**Late submission till 7 days**	**Submission on time**	
	1 points	2 points	3 points	

Faculty Incharge : Ms. Sana Shaikh