

Aim: To Install Cloudera Quickstart VM on VirtualBox

Cloudera is a software that provides a platform for data analytics, data warehousing, and machine learning. Initially, Cloudera started as an open-source Apache Hadoop distribution project, commonly known as Cloudera Distribution for Hadoop or CDH. It contains Apache Hadoop and other related projects where all the components are 100% open-source under Apache License.

Cloudera provides virtual machine images of complete Apache Hadoop clusters, making it easy to get started with Cloudera CDH. The following topics will be covered in this assignment on Cloudera QuickStart VM Installation.

1. What is Cloudera QuickStart VM?
2. Cloudera QuickStart VM Installation - Prerequisites
3. Downloading the Cloudera QuickStart VM
4. Cloudera QuickStart VM Installation on windows

What Is Cloudera QuickStart VM?

Cloudera QuickStart VM includes everything that you would need for using CDH, Impala, Cloudera Search, and Cloudera Manager. The Cloudera QuickStart VM uses a package-based install that allows you to work with or without the Cloudera Manager. It has a sample of Cloudera's platform for "Big Data."

Cloudera QuickStart VM Installation - Prerequisites

A virtual machine such as Oracle Virtual Box or VMWare RAM of 12+ GB. That is 4+ GB for the operating system and 8+ GB for Cloudera

80GB hard disk

Download Oracle Virtual Box from <https://www.virtualbox.org/wiki/Downloads> and install it in your system



Link: <https://www.cloudera.com/downloads.html>

Cloudera CDH
Cloudera's open source software distribution including Apache Hadoop and additional key open source projects
[Download CDH >](#)
[Download Phoenix for CDH >](#)

Hortonworks Data Platform (HDP)
Hortonworks Data Platform (HDP) helps enterprises gain insights from structured and unstructured data. It is an open source framework for distributed storage and processing of large, multi-source data sets.
[Download the Hortonworks Data Platform \(HDP\) >](#)
[Legacy HDP releases >](#)

Cloudera Workload XM
Workload XM proactively assists, de-risks, and advises
[Download now >](#)

Cloudera DataFlow (Ambari)
Cloudera DataFlow (Ambari)—formerly Hortonworks

If already an account, then login else first register the account

Please Log In

Email Address*

Password*

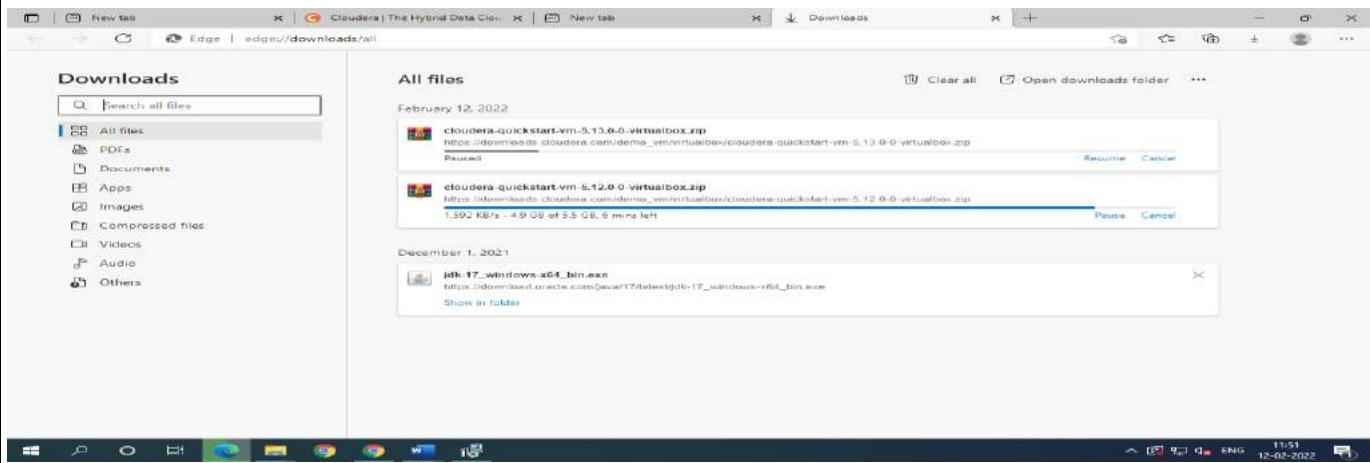
Log In

Cloudera uses cookies to provide and improve our site's services. By using this site, you consent to use of cookies as outlined in [Cloudera's Privacy and Data Policies](#).

Accept Cookies

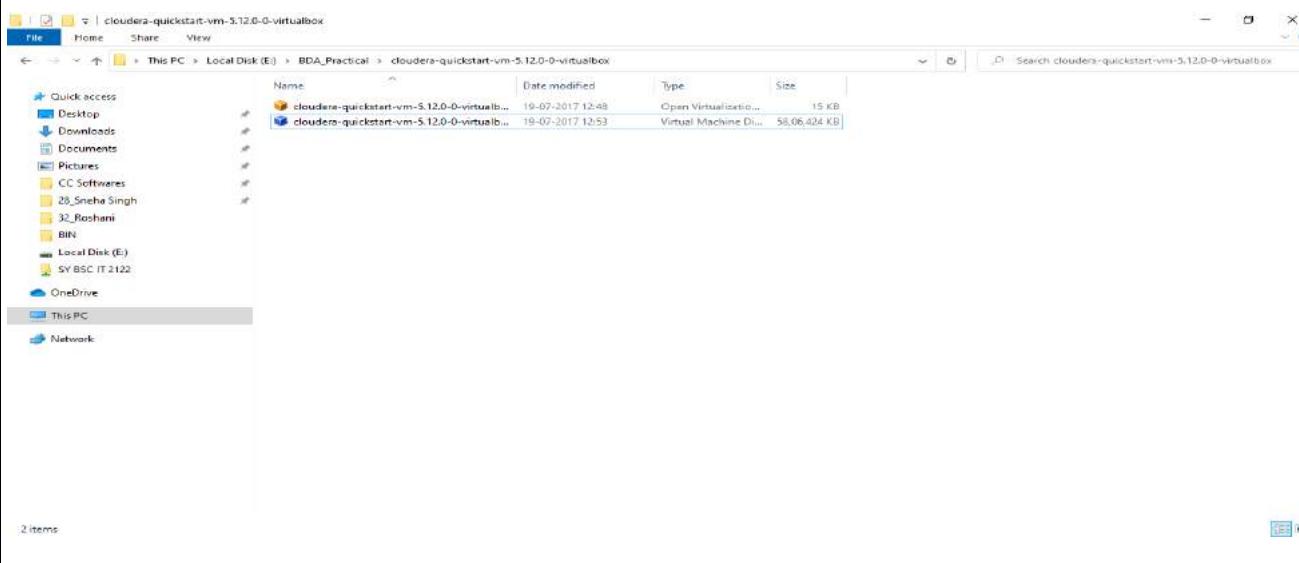
cloudera download link –

https://downloads.cloudera.com/demo_vm/virtualbox/cloudera-quickstart-vm-5.13.0-0-virtualbox.zip



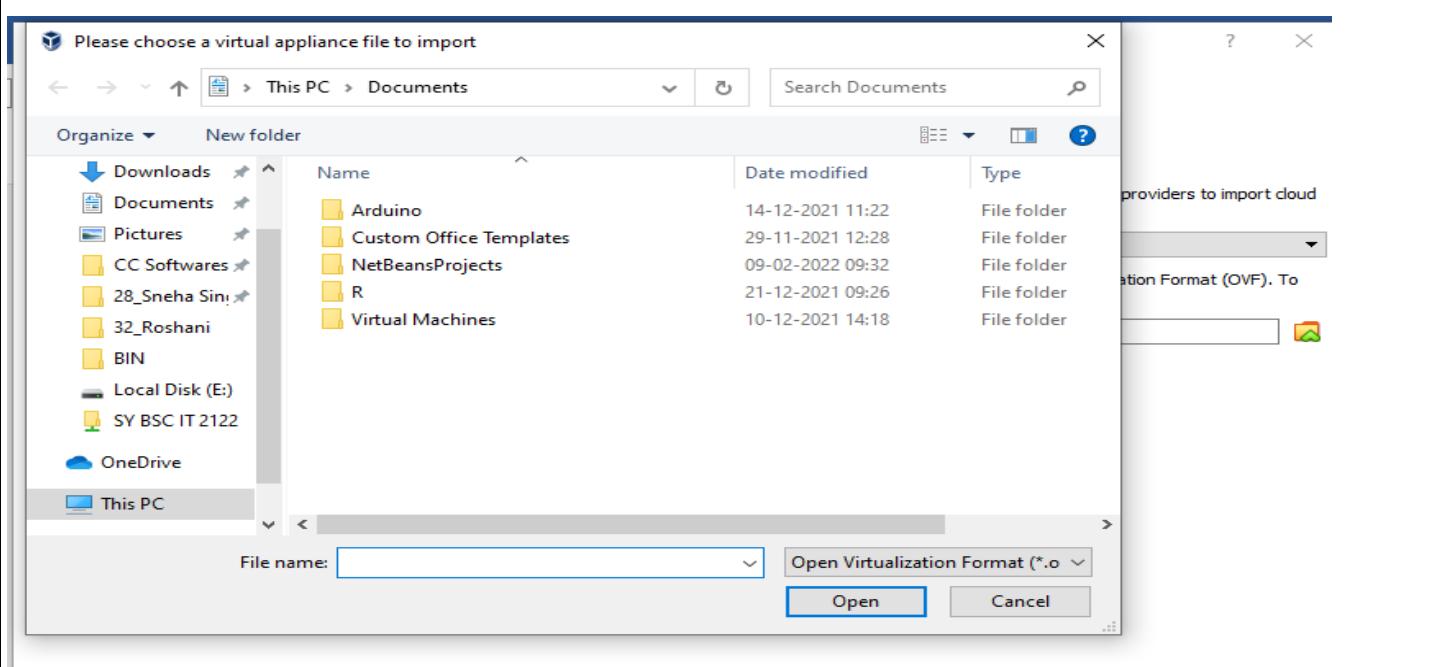
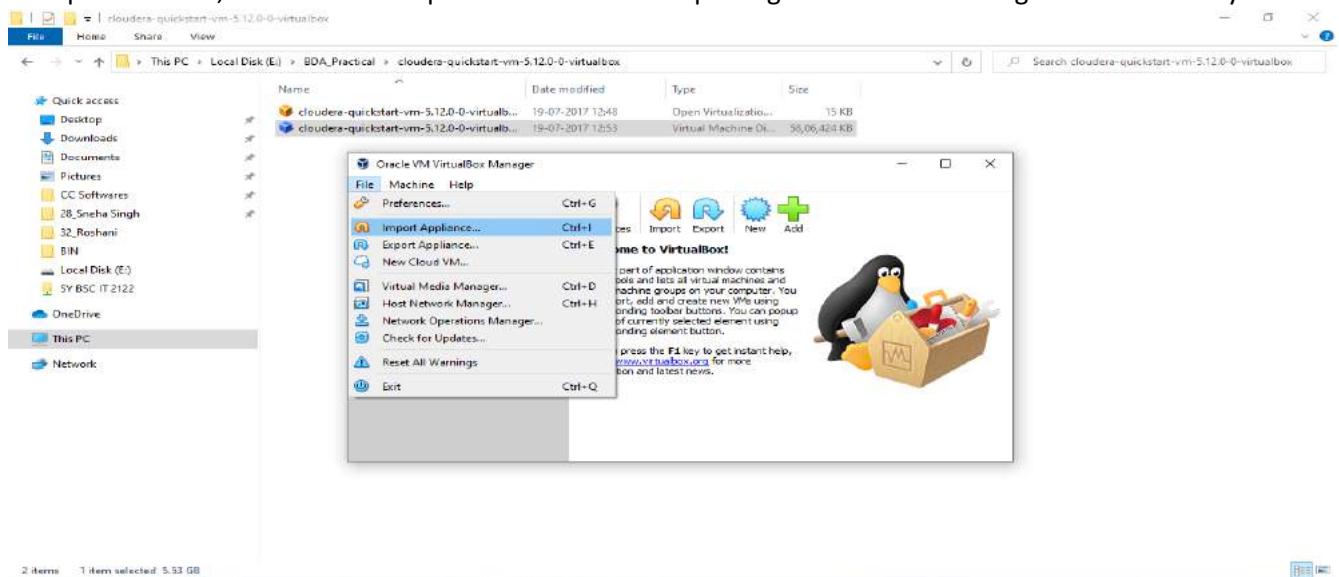
Downloading the Cloudera QuickStart VM

- The Cloudera QuickStart VMs are openly available as Zip archives in VirtualBox, VMware and KVM formats. To download the VM, search for <https://www.cloudera.com/downloads.html> and select the appropriate version of CDH that you require.
- Click on the 'GET IT NOW' button, and it will prompt you to fill in your details.
- Once the file is downloaded, go to the download folder and unzip these files. It can then be used to set up a single node Cloudera cluster.
- Shown below are the two virtual images of Cloudera QuickStart VM.
- Now that the downloading process is done with, let's move forward with this Cloudera QuickStart VM Installation guide and see the actual process.



Cloudera QuickStart VM Installation

- Before setting up the Cloudera Virtual Machine, you would need to have a virtual machine such as VMware or Oracle VirtualBox on your system.
- In this case, we are using Oracle VirtualBox to set up the Cloudera QuickStart VM.
- In order to download and install the Oracle VirtualBox on your operating system, click on the following link: Oracle VirtualBox(<https://www.virtualbox.org/wiki/Downloads>).
- To set up the Cloudera QuickStart VM in your Oracle VirtualBox Manager, click on 'File' and then select 'Import Appliance'.
- Choose the QuickStart VM image by looking into your downloads. Click on 'Open' and then 'Next'. Now you can see the specifications, then click on 'Import'. This will start importing the virtual disk image .vmdk file into your VM box.



- Click on "Open" and Wait for a while, as the importing finishes.
- The next step is to go ahead and set up a Cloudera QuickStart VM for practice. Once the importing is complete, you can see the Cloudera QuickStart VM on the left side panel.

Name: Pavan Yadav

Practical 1

Please choose a virtual appliance file to import

Organize New folder

Name Date modified Type

cloudera-quickstart-vm-5.12.0-0-virtualb... 19-07-2017 12:48 Open Virtualizat...

Downloads Documents Pictures CC Softwares 28_Sneha Sini 32_Roshani BIN Local Disk (E:) SY BSC IT 2122 OneDrive This PC

File name: cloudera-quickstart-vm-5.12.0-0-virtualbox Open Virtualization Format (*.o...

Open Cancel

Import Virtual Appliance

Appliance to import

Please choose the source to import appliance from. This can be a local file system to import OVF archive or one of known cloud service providers to import cloud VM from.

Source: Local File System

Please choose a file to import the virtual appliance from. VirtualBox currently supports importing appliances saved in the Open Virtualization Format (OVF). To continue, select the file to import below.

File: E:\BDA_Practical\cloudera-quickstart-vm-5.12.0-0-virtualbox\cloudera-quickstart-vm-5.12.0-0-virtualbox.ovf

Appliance settings

These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the properties shown by double-clicking on the items and disable others using the check boxes below.

Virtual System 1	
	Name: cloudera-quickstart-vm-5.12.0-0-virtualbox
	Guest OS Type: Red Hat (64-bit)
	CPU: 1
	RAM: 4096 MB
	DVD: <input checked="" type="checkbox"/>
	Network Adapter: Intel PRO/1000 MT Desktop (82540EM)
	Storage Controller (IDE): PIIX4
	Virtual Disk Image: cloudera-quickstart-vm-5.12.0-0-virtualbox-disk1.vmdk
	Base Folder: C:\Users\Admin\VirtualBox VMs
	Primary Group: /

Machine Base Folder: C:\Users\Admin\VirtualBox VMs

MAC Address Policy: Include only NAT network adapter MAC addresses

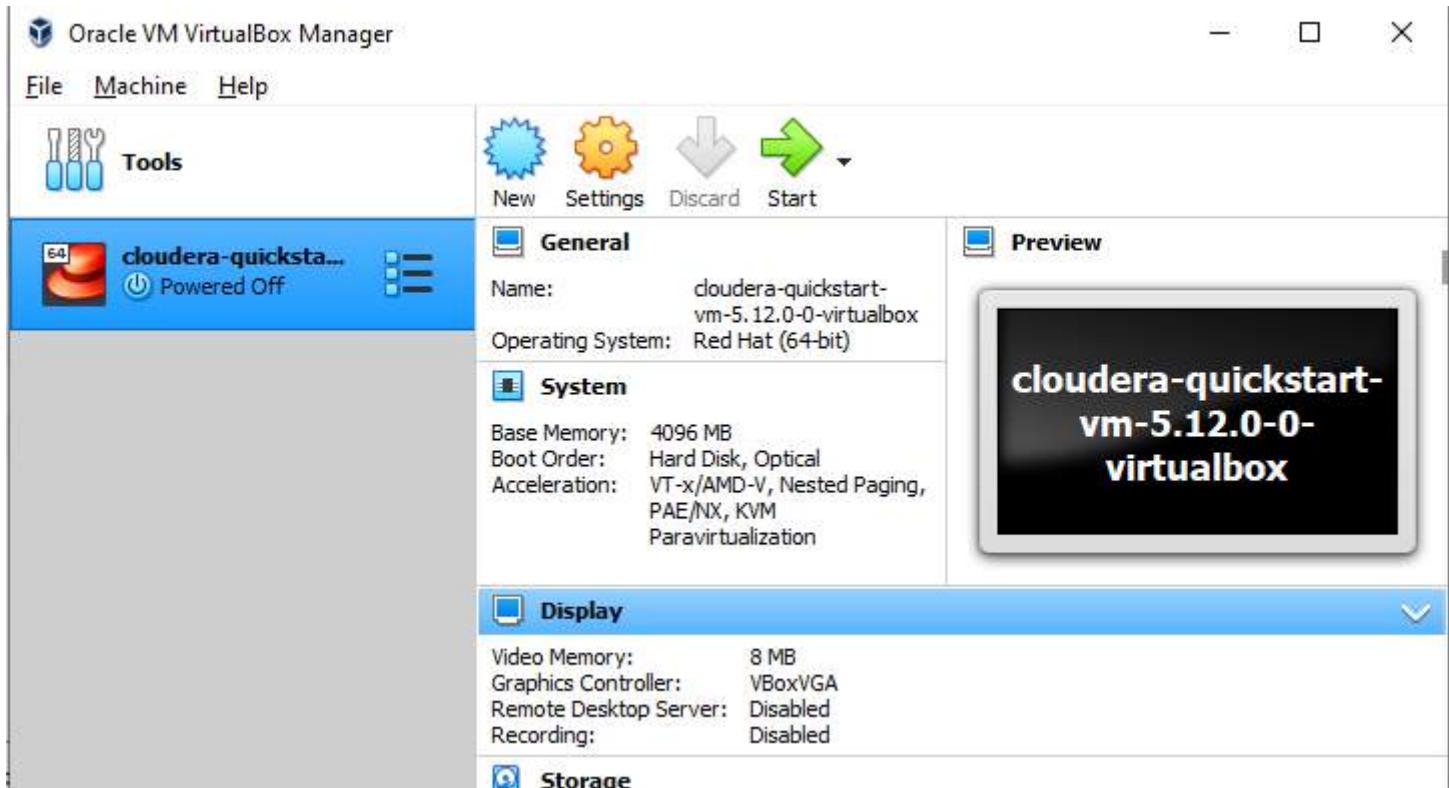
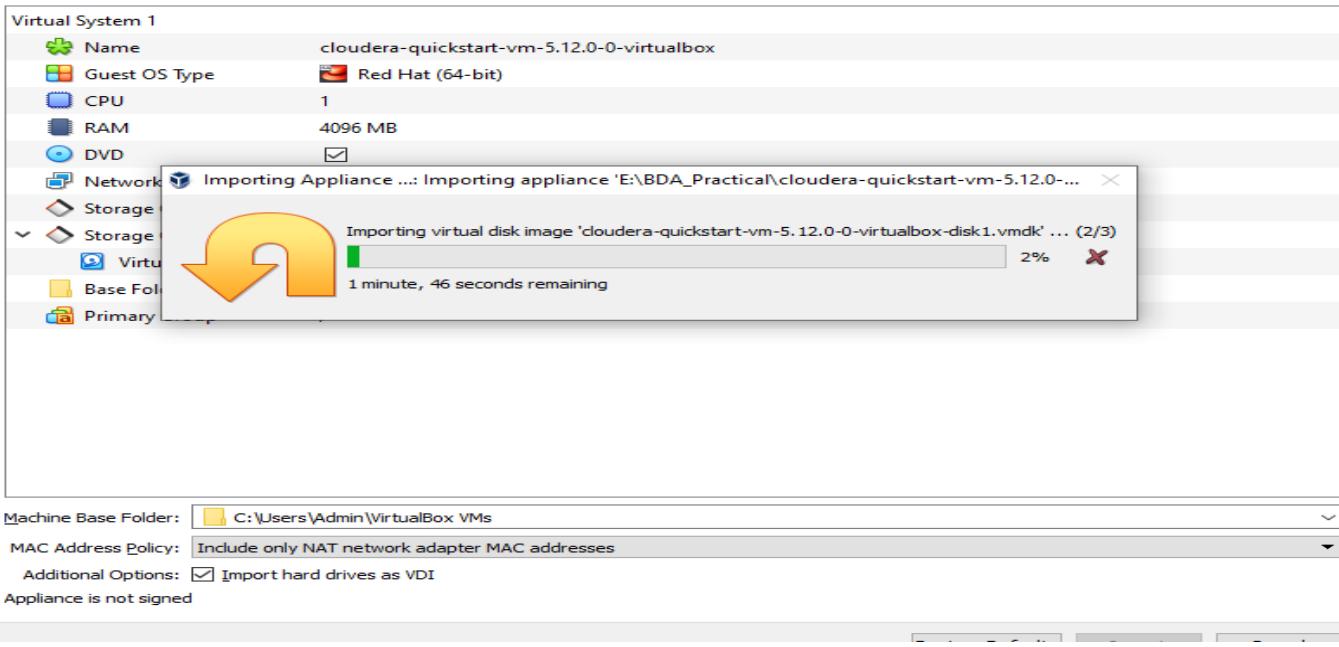
Additional Options: Import hard drives as VDI

Appliance is not signed

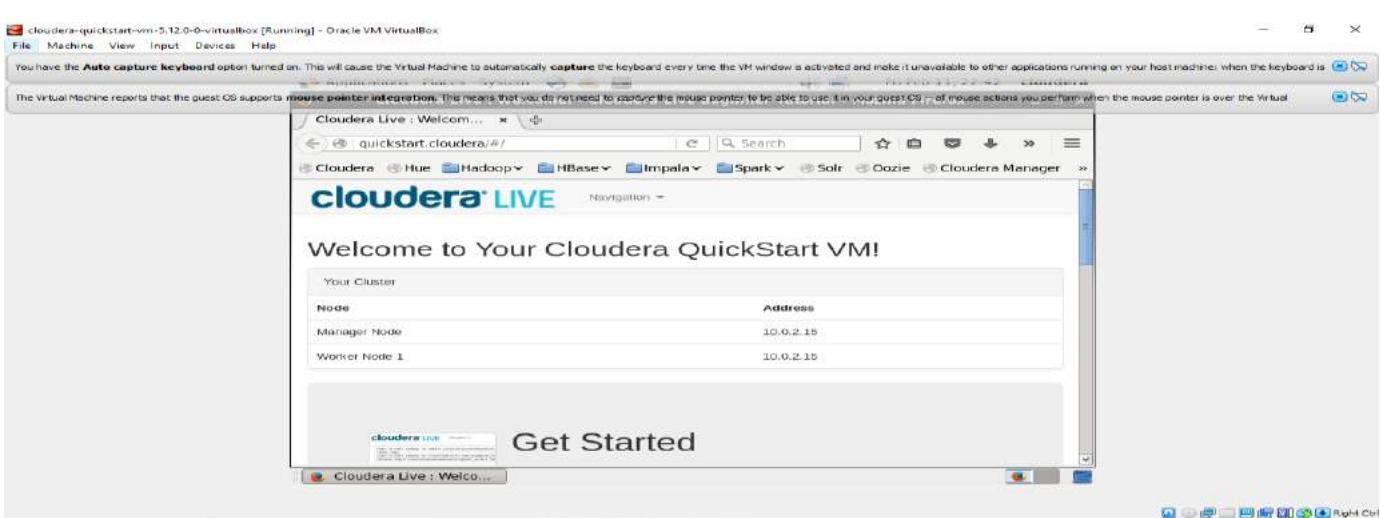
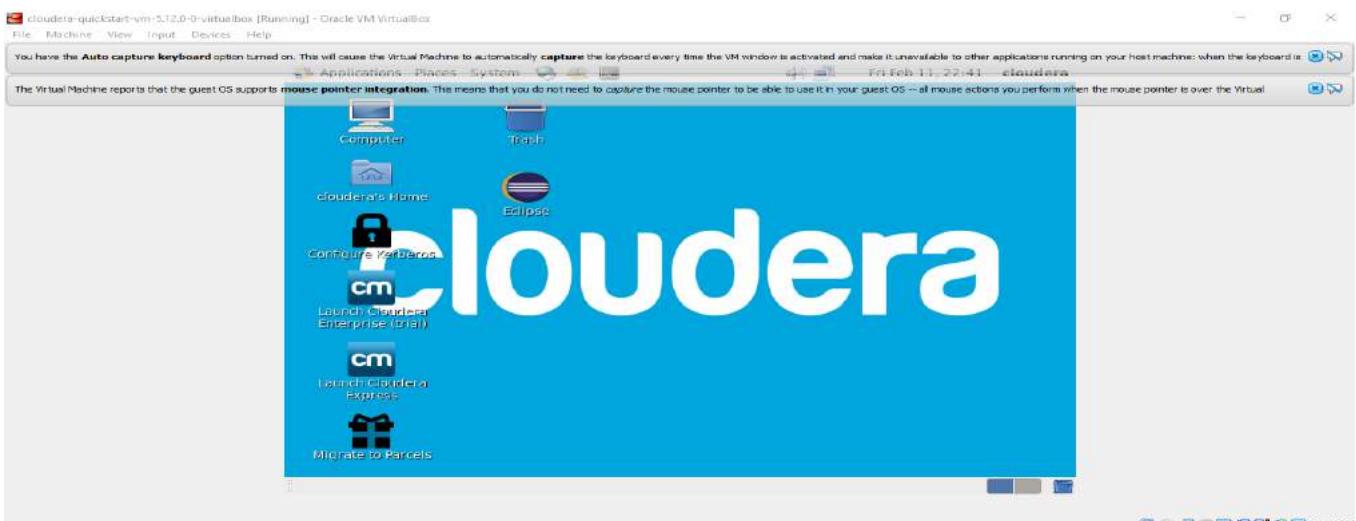
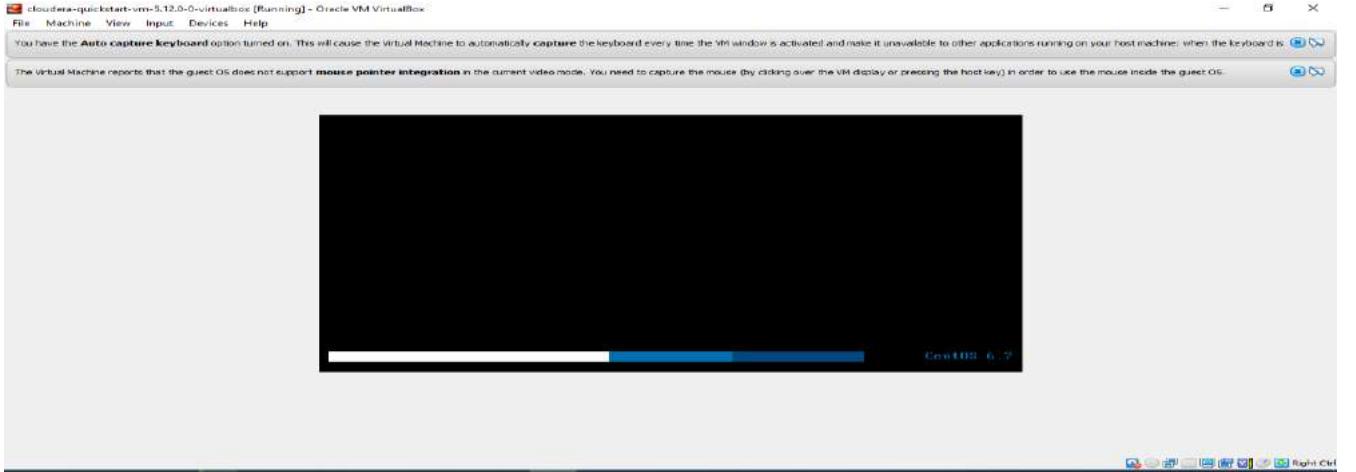
← Import Virtual Appliance

Appliance settings

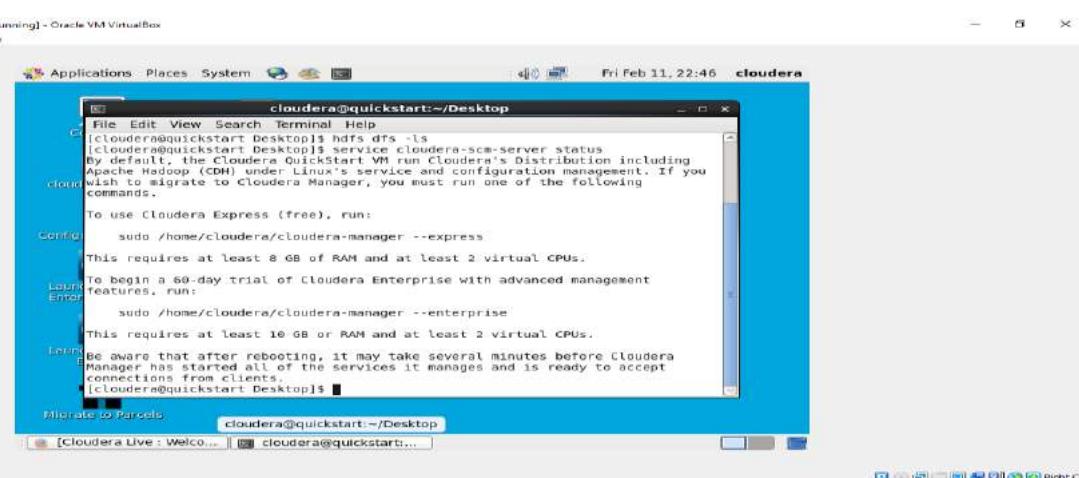
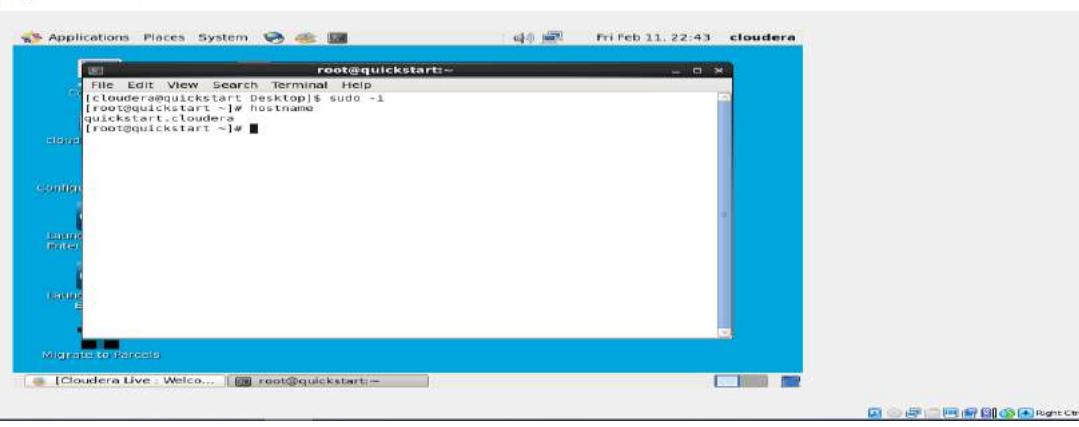
These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the properties shown by double-clicking on the items and disable others using the check boxes below.



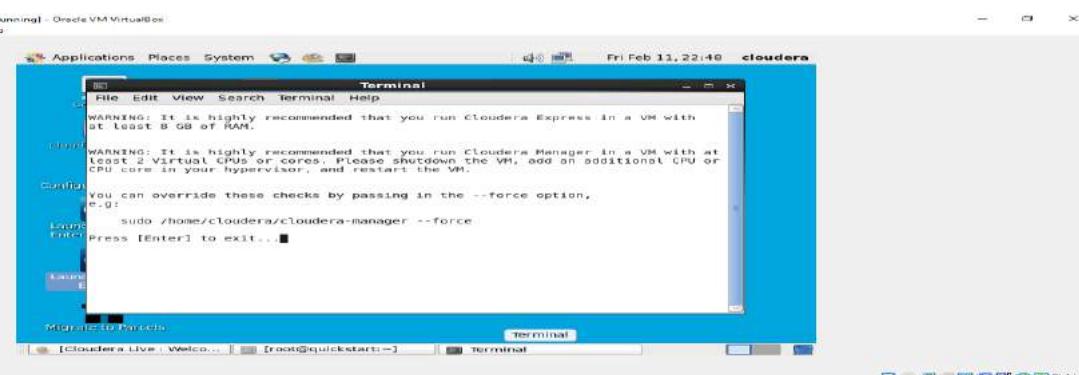
- The next step will be going ahead and starting the machine by clicking the 'Start' symbol on top.
- Once your machine comes on, it will look like this:



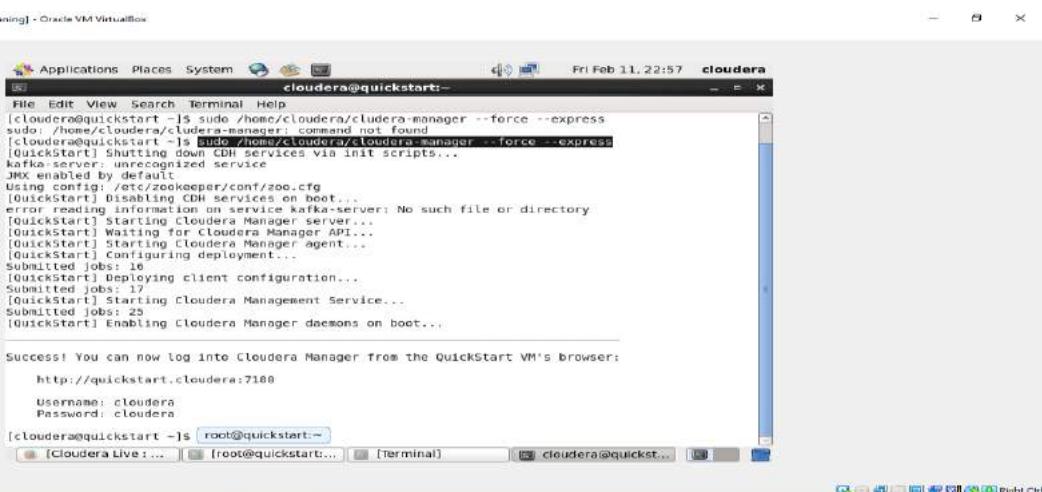
- Next, we have to follow a few steps to gain admin console access. You need to click on the terminal present on top of the desktop screen, and type in the following:
 1. **hostname** # This shows the hostname which will be quickstart.cloudera
 2. **hdfs dfs -ls /** # Checks if you have access and if your cluster is working. It displays what exists on your HDFS location by default
 3. **service cloudera-scm-server status** # Tells what command you have to type to use cloudera express free
 4. **su -** #Login as root
 5. **service cloudera-scm-server status** # The password for root is cloudera
- Once you see that your HDFS access is working fine, you can close the terminal. Then, you have to click on the following icon that says 'Launch Cloudera Express'.



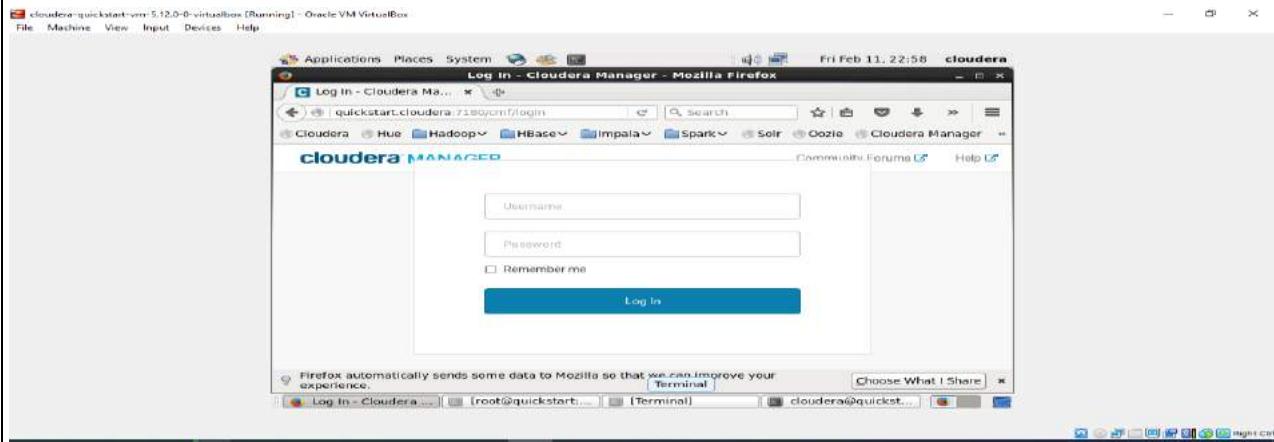
- You are required to copy the command, and run it on a separate terminal. Hence, open a new terminal, and use the below command to close the Cloudera based services. It will restart the services, after which you can access your admin console.



sudo /home/cloudera/cloudera-manager --force --express



- Now that our deployment has been configured, client configurations have also been deployed. Additionally, it has restarted the Cloudera Management Service, which gives access to the Cloudera QuickStart admin console with the help of a username and password.
- Go on and open up the browser and change the port number to 7180.
- You can log in to the Cloudera Manager by providing your username and password.



- You can go ahead and restart the services now. It will ensure that the cluster becomes accessible either by Hue as a web interface or Cloudera QuickStart Terminal, where you can write your commands.

Aim: To implement Various Hadoop HDFS Commands

What is Hadoop?

Apache Hadoop is an open source software framework used to develop data processing applications which are executed in a distributed computing environment.

Applications built using HADOOP are run on large data sets distributed across clusters of commodity computers. Commodity computers are cheap and widely available. These are mainly useful for achieving greater computational power at low cost.

Similar to data residing in a local file system of a personal computer system, in Hadoop, data resides in a distributed file system which is called as a Hadoop Distributed File system. The processing model is based on 'Data Locality' concept wherein computational logic is sent to cluster nodes (server) containing data. This computational logic is nothing, but a compiled version of a program written in a high-level language such as Java. Such a program, processes data stored in Hadoop HDFS.

Hadoop Distributed File System (HDFS) - Data Storage and Management

This is the most important component of the Hadoop ecosystem. HDFS is Hadoop's primary storage system. Hadoop Distributed File System (HDFS) is a Java-based file system that provides reliable, fault tolerance and accessible data storage for the big data. HDFS is a distributed file system that runs on conventional hardware. HDFS is already configured with the default settings for many installations. Typically, a large cluster configuration is required. Hadoop interacts directly with HDFS using commands. When comes to HDFS, there are also two components can be identified, which are known as Name Node and Data Node.

Name Node

It is also known as Master node. Here, it does not store actual data or datasets. Name Node stores the Meta data, for an example, the number of calls transform from a tower, their position, where the end users are getting the call, the Data node data and other details. Basically, this contains files and directories. The tasks of Name node can be recognized as follows.

- Managing file system namespace
- Controlling the access of clients to files
- Executing file system through naming, opening, closing files and directories

Data Node

Data node is called as Slave. Data node is responsible for the effective storage of data in HDFS. The data node completes read and write operations on customer request. They also send signals, known as heartbeats, to the name node. These heartbeats show the status of the data node. Replica Block of Data node consists of two files in the file system. The first file is for data and the second for registry metadata. HDFS metadata contains a data control. At startup, each Data node is connected to the appropriate Name node and grasp. The ID of the Data Node namespace and the software version are controlled by the handshake. If a discrepancy is detected, Data Node is automatically disabled. When comes to tasks of Data node, those can be detailed as follows.

- This is consisting of operations like block replica creation, deletion, and replication according to the instruction of Name node
- Managing data storage of the system

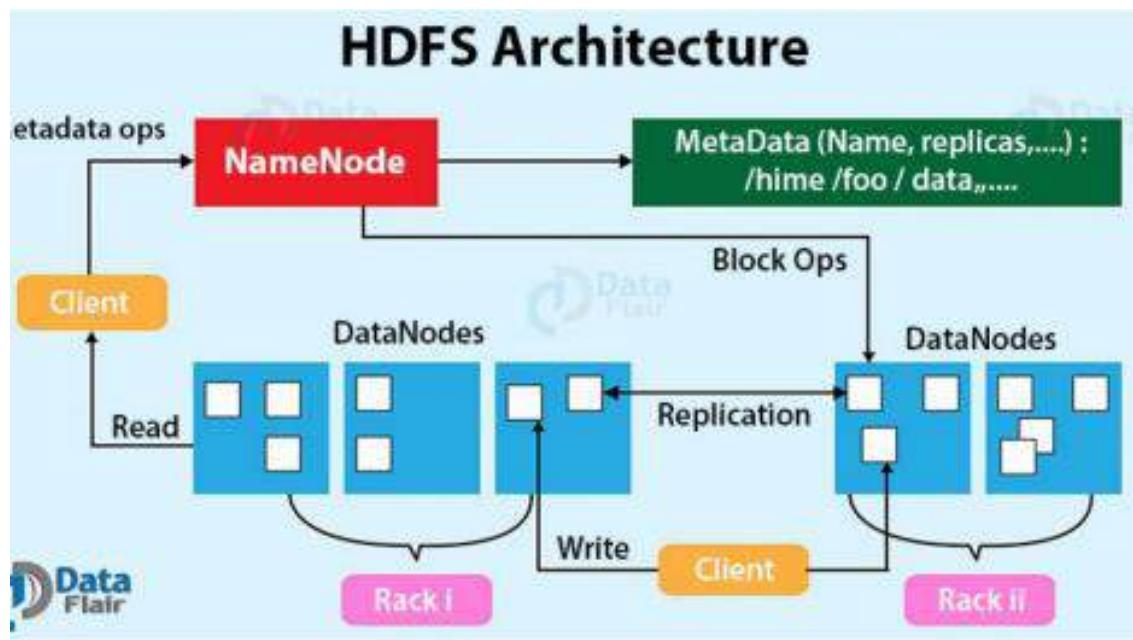
Processing and Computation – Hadoop MapReduce

When comes to Hadoop MapReduce, that is the main component of the Hadoop, that provides data processing. MapReduce is can be identified as an easy-to-write application framework that processes the large amount of structured and unstructured data stored in the Hadoop distributed file system.

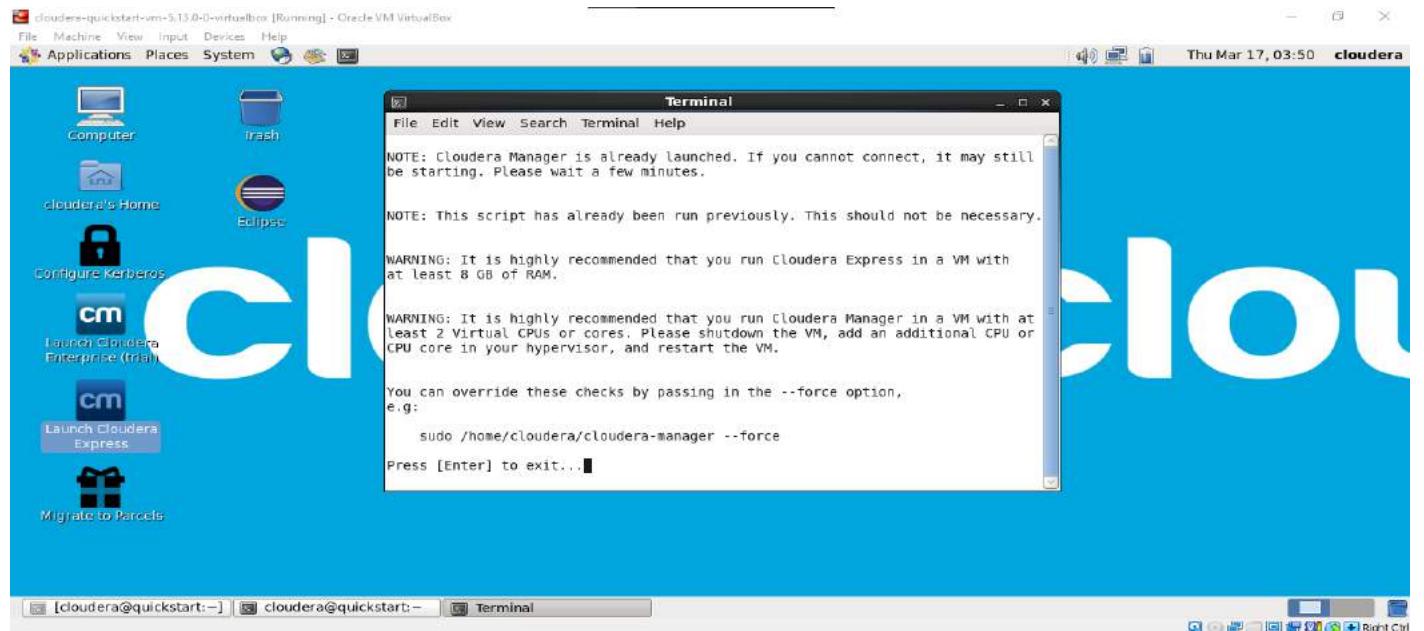
MapReduce programs are parallel, so they are very useful for large-scale data analysis using multiple clusters. Therefore, this parallelism increases the speed and reliability of the cluster. In MapReduce, there are two functions, Map function and Reduce function.

Two functions can be identified, map function and reduce function.

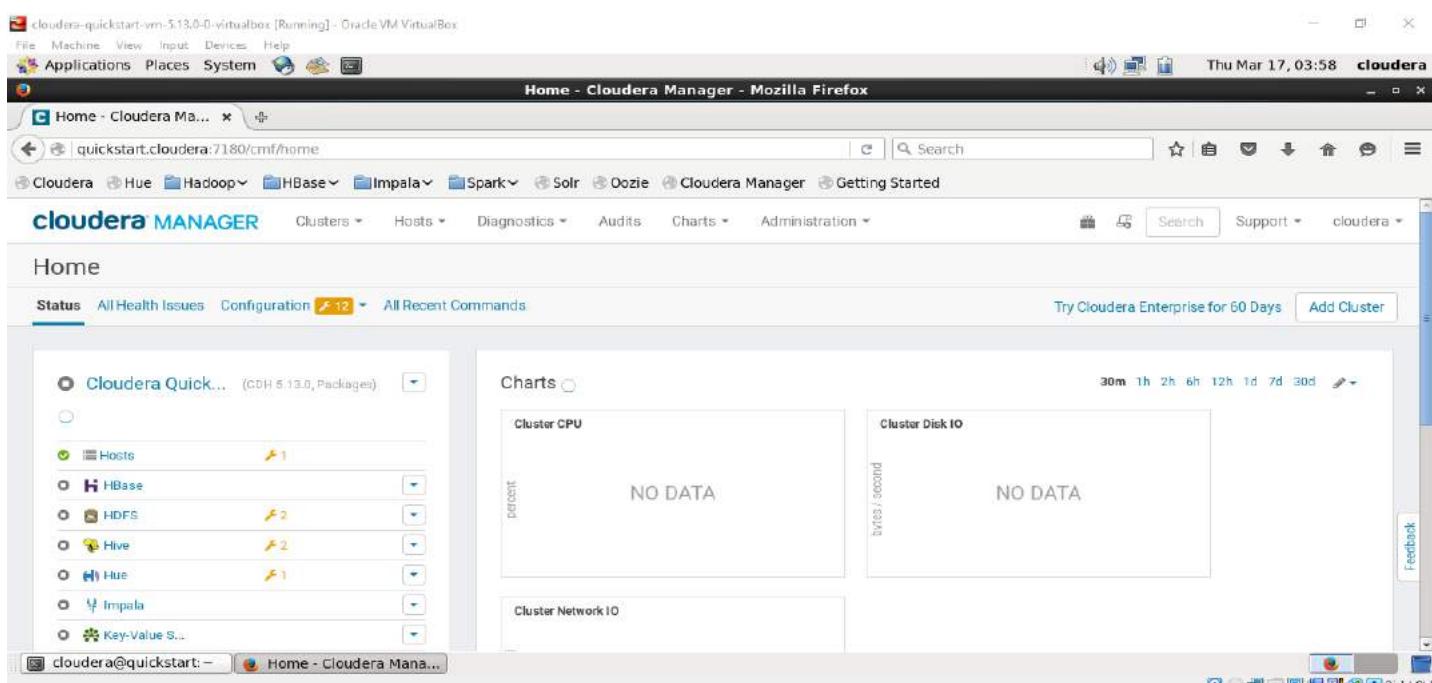
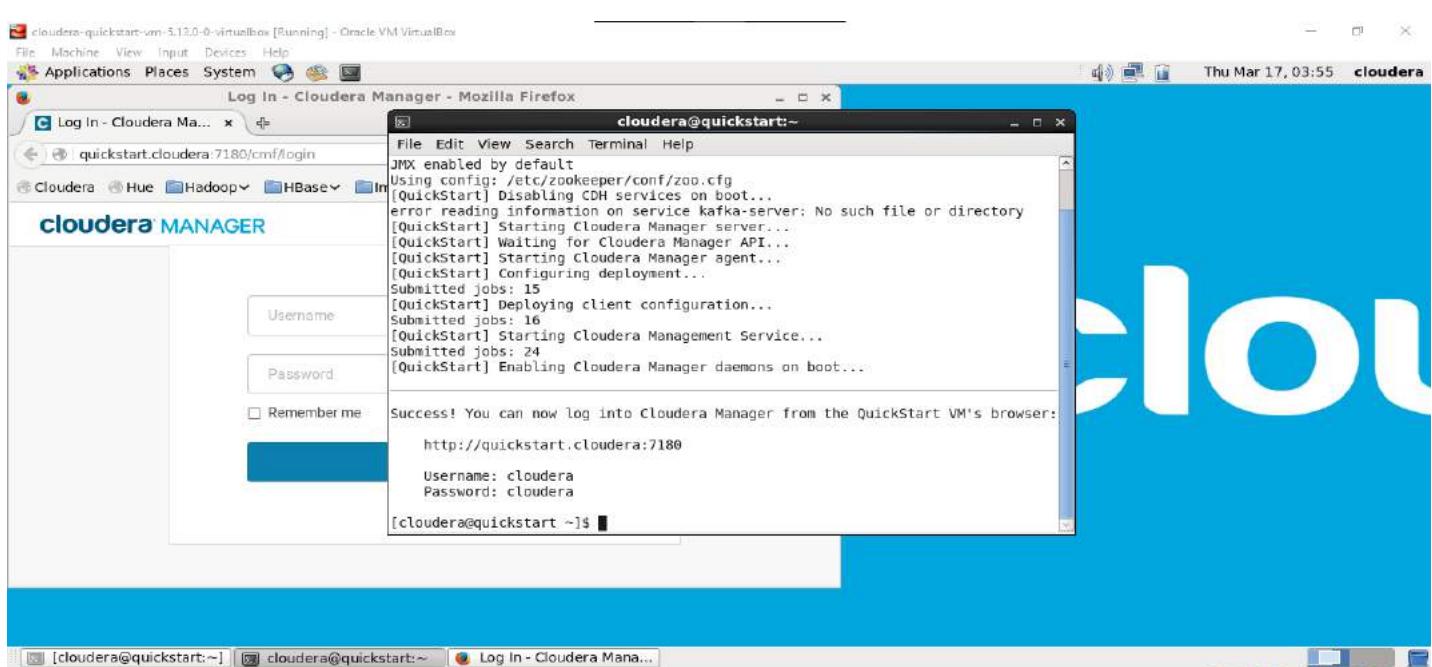
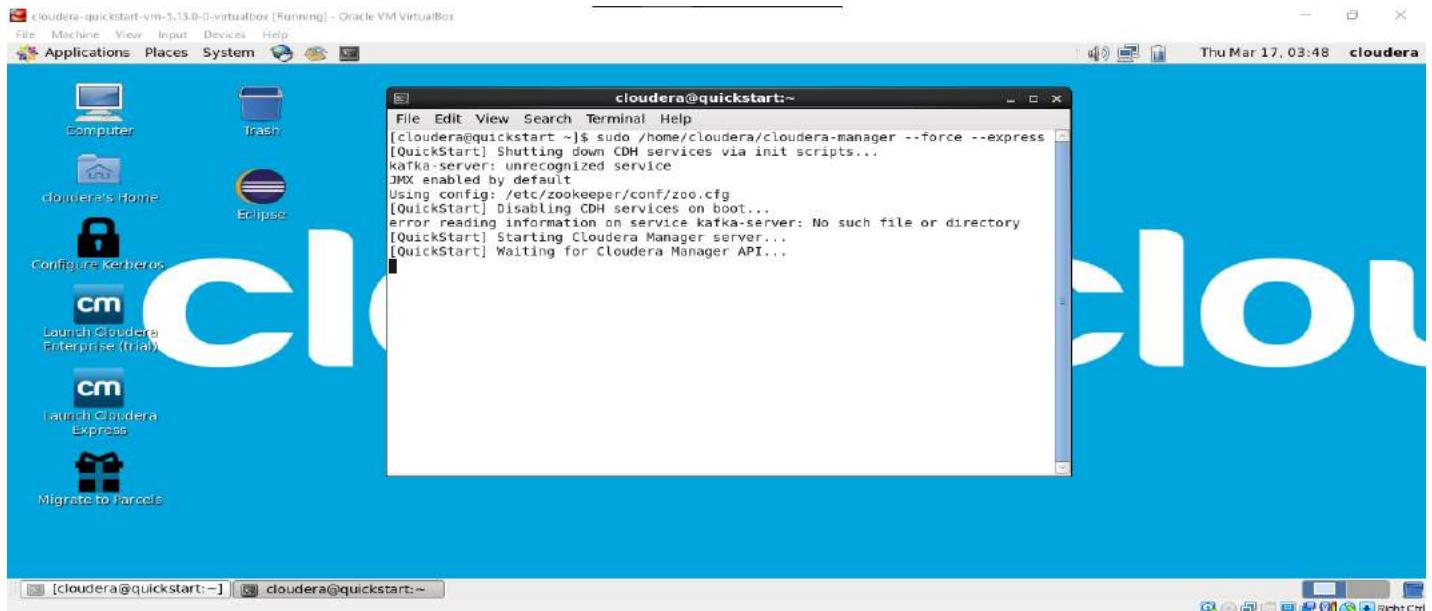
- The map function retrieves a data set and converts it to another data set. Each element is divided into processing (key / value pairs).
- The Reduce function accepts the Map output as an input and integrates these data nodes based on the key and changes the key value accordingly.



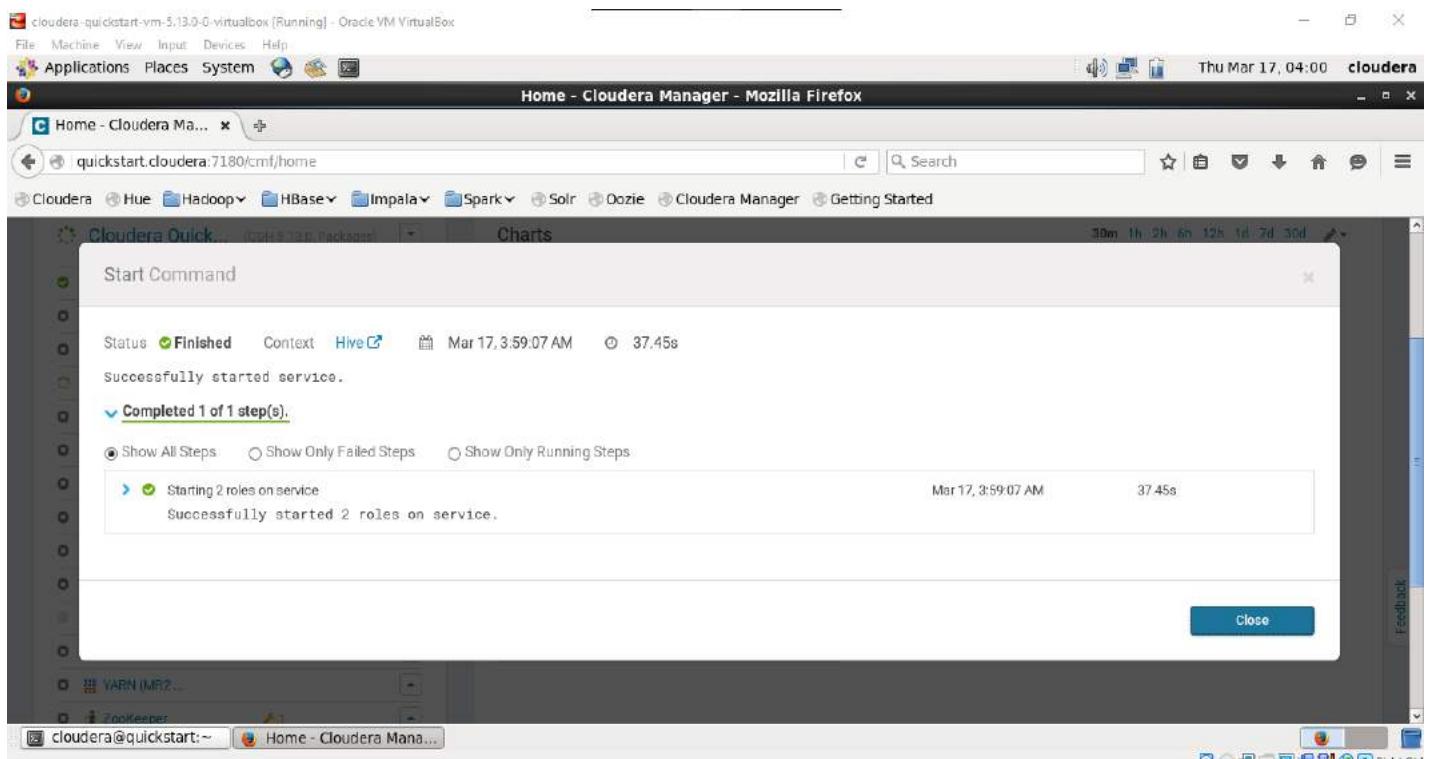
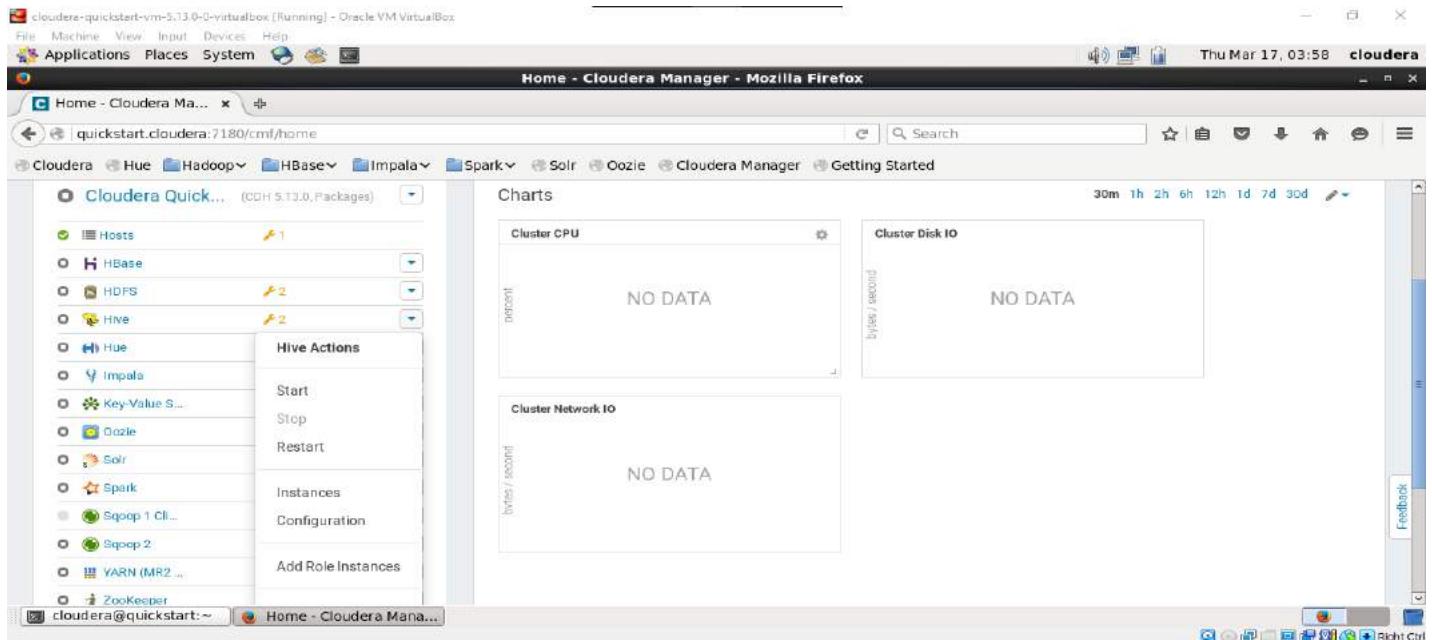
Copy the command from the Launch Cloudera Express



Type **sudo /home/cloudera/cloudera-manager --force --express** in terminal



Click on down arrow beside HDFS and select start



Hadoop HDFS Commands

First go to the **cd Desktop/**

1. `hadoop version` is used to check version of Hadoop system

```
[cloudera@quickstart ~]$ hadoop version
Hadoop 2.6.0-cdh5.13.0
subversion http://github.com/cloudera/hadoop -r 42e8860b182e55321bd5f5605264da4a
compiled by jenkins on 2017-10-04T18:08Z
Compiled with protoc 2.5.0
From source with checksum 5e84c185f8a22158e2b0e4b8f85311
This command was run using /usr/lib/hadoop/hadoop-common-2.6.0-cdh5.13.0.jar
Configure [cloudera@quickstart ~]$
```

2. `hdfs dfs -ls /` lists all the files and folders present in HDFS location. It lists the contents of the directory specified by path, names, permissions, owner, size and modification date for each entry. `hdfs dfs` is the command specific to HDFS.

```
[cloudera@quickstart ~]$ cd Desktop/
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx  - hdfs  supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x  - hbase  supergroup          0 2022-03-17 03:06 /hbase
drwxr-xr-x  - solr   supergroup          0 2017-10-23 09:18 /solr
drwxrwxrwt  - hdfs  supergroup          0 2022-03-17 00:48 /tmp
drwxr-xr-x  - hdfs  supergroup          0 2022-03-17 07:26 /user
drwxr-xr-x  - hdfs  supergroup          0 2017-10-23 09:17 /var
[cloudera@quickstart Desktop]$
```

3. If getting any error on permission access use `export HADOOP_USER_NAME=hdfs` after this command run previous command again and use `hdfs dfs -ls /` to check if new directory "rjcloca" is created.

```
ls: Failed to connect to quickstart.cloudera:8020 from quickstart.cloudera:10.0.2.15: Connection refused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
[cloudera@quickstart ~]$ export HADOOP_USER_NAME=hdfs
[cloudera@quickstart ~]$
```

Restart then execute the `hdfs dfs -ls /`

4. If getting error on safe mode then run `hadoop dfadmin -safemode leave`

```
[cloudera@quickstart ~]$ export HADOOP_USER_NAME=hdfs
[cloudera@quickstart ~]$ hadoop dfadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Safe mode is OFF
[cloudera@quickstart ~]$
```

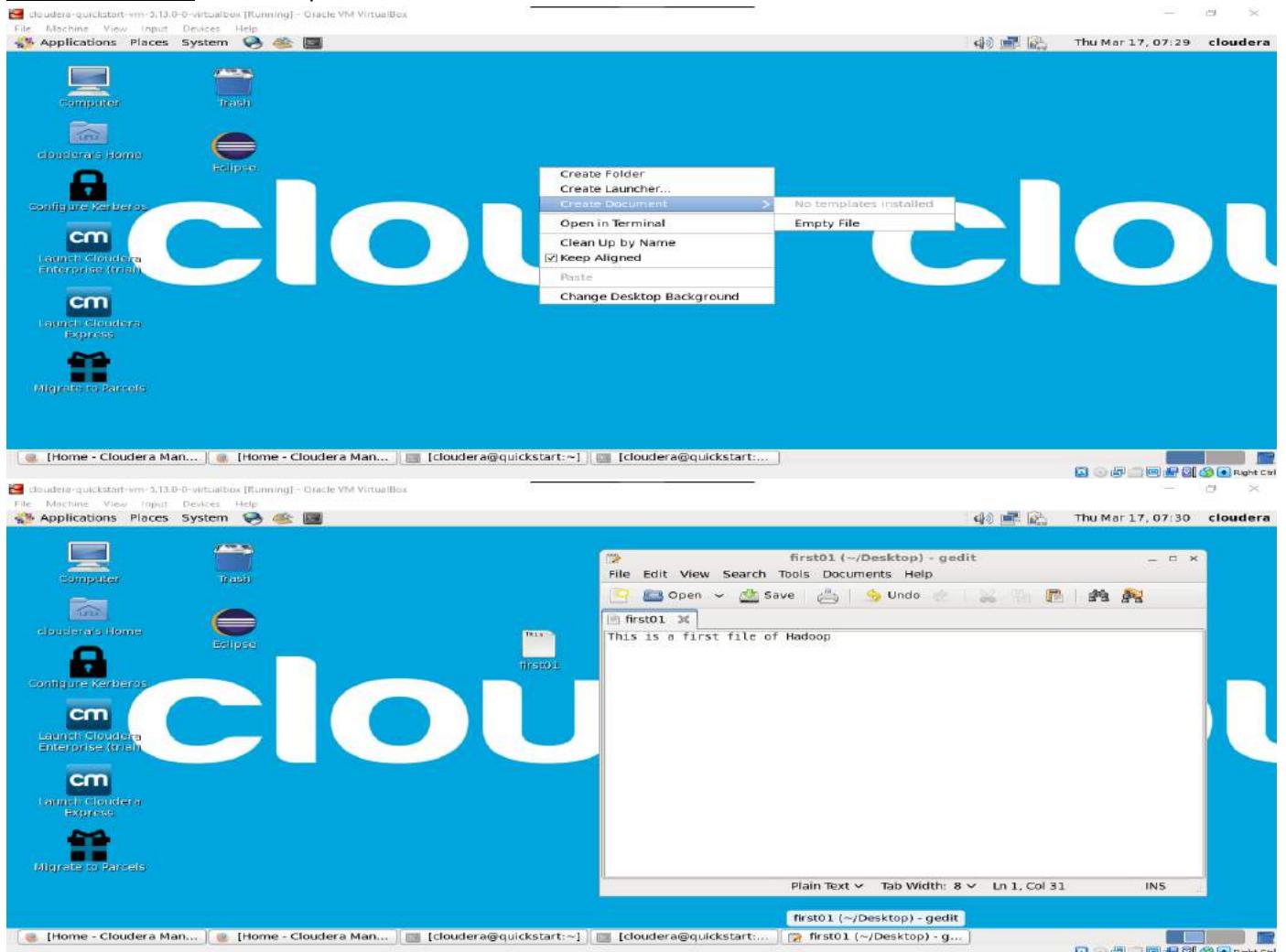
5. `hadoop fs -ls` / displays a list of content of a directory specified in the path provided by the user.

```
[cloudera@quickstart ~]$ cd Desktop/
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx  - hdfs  supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x  - hbase  supergroup          0 2022-03-17 03:06 /hbase
drwxr-xr-x  - solr   solr              0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs  supergroup          0 2022-03-17 00:46 /tmp
drwxr-xr-x  - hdfs  supergroup          0 2022-03-17 07:26 /user
drwxr-xr-x  - hdfs  supergroup          0 2017-10-23 09:17 /var
[cloudera@quickstart Desktop]$ hadoop fs -ls /
Found 6 items
drwxrwxrwx  - hdfs  supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x  - hbase  supergroup          0 2022-03-17 03:06 /hbase
drwxr-xr-x  - solr   solr              0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs  supergroup          0 2022-03-17 00:46 /tmp
drwxr-xr-x  - hdfs  supergroup          0 2022-03-17 07:26 /user
drwxr-xr-x  - hdfs  supergroup          0 2017-10-23 09:17 /var
```

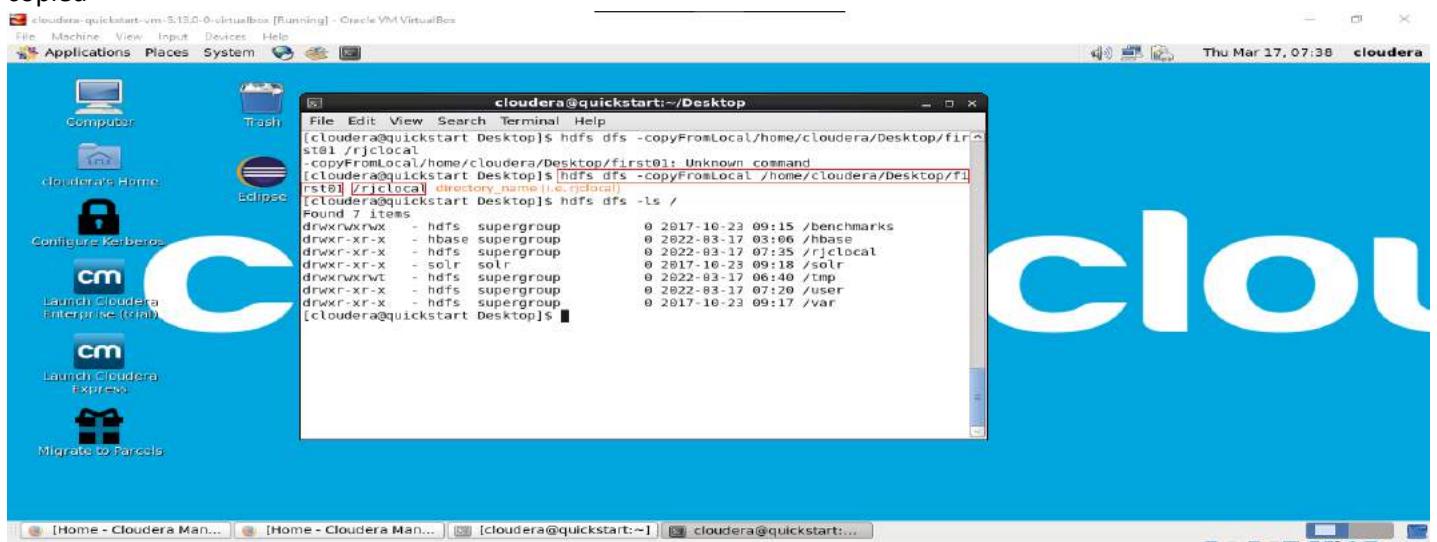
6. `hdfs dfs -mkdir /rjlocal`

```
[cloudera@quickstart ~]$ cd Desktop/
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx  - hbase  supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x  - solr   solr              0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs  supergroup          0 2022-03-17 00:46 /tmp
drwxr-xr-x  - hdfs  supergroup          0 2022-03-17 07:26 /user
drwxr-xr-x  - hdfs  supergroup          0 2017-10-23 09:17 /var
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx  - hdfs  supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x  - hbase  supergroup          0 2022-03-17 03:06 /hbase
drwxr-xr-x  - solr   solr              0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs  supergroup          0 2022-03-17 00:46 /tmp
drwxr-xr-x  - hdfs  supergroup          0 2022-03-17 07:26 /user
drwxr-xr-x  - hdfs  supergroup          0 2017-10-23 09:17 /var
[cloudera@quickstart Desktop]$ hdfs dfs -mkdir /rjlocal
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 7 items
drwxrwxrwx  - hdfs  supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x  - hbase  supergroup          0 2022-03-17 03:06 /hbase
drwxr-xr-x  - hdfs  supergroup          0 2022-03-17 07:27 /rjlocal
drwxr-xr-x  - solr   solr              0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs  supergroup          0 2022-03-17 00:46 /tmp
drwxr-xr-x  - hdfs  supergroup          0 2022-03-17 07:26 /user
drwxr-xr-x  - hdfs  supergroup          0 2017-10-23 09:17 /var
```

7. Create a new file on desktop



8. Use `hdfs dfs -copyFromLocal /home/cloudera/Desktop/<file_name> /<directory_name>`- this command copies <file_name> file to HDFS <directory_name> directory and use `hdfs dfs -ls /rjclocal` to check if file is successfully copied



```

cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
[cloudera@quickstart Desktop]$ hdfs dfs -copyFromLocal/home/cloudera/Desktop/first01 /rjlocal
[cloudera@quickstart Desktop]$ hdfs dfs -copyFromLocal/home/cloudera/Desktop/first01: Unknown command
[cloudera@quickstart Desktop]$ hdfs dfs -copyFromLocal /home/cloudera/Desktop/first01 /rjlocal
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 7 items
drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup 0 2022-03-17 03:06 /hbase
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:35 /rjlocal
drwxr-xr-x - solr supergroup 0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup 0 2022-03-17 06:48 /tmp
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:28 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /var
[cloudera@quickstart Desktop]$ hdfs dfs -ls /rjlocal
Found 1 items
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 07:35 /rjlocal/first01
[cloudera@quickstart Desktop]$

```

9. `hdfs dfs -cat <directory_name>/<file_name>` reads content of file

```

On Hadoop File system file content copied from the local
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
[cloudera@quickstart Desktop]$ hdfs dfs -cat /rjlocal/first01
This is a first file of Hadoop
[cloudera@quickstart Desktop]$

```

A gedit window titled "first01 (~/Desktop) - gedit" shows the text "This is a first file of Hadoop".

10. First we create a new directory in HDFS “`newdir`” `hdfs dfs -mkdir <dir_name>` and then use `hdfs dfs -cp <dir1_name>/<dir_file_name> <dir2_name>` then check if file is copied successfully by using `hdfs dfs -ls <dir2_name>` `cp` command is used to copy file from one directory to another within HDFS

Make directory

```

cloudera@quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
[cloudera@quickstart Desktop]$ hdfs dfs -cat /rjlocal/first01
This is a first file of Hadoop
[cloudera@quickstart Desktop]$ hdfs dfs -mkdir /newdir
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 8 items
drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup 0 2022-03-17 03:06 /hbase
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:53 /newdir
drwxr-xr-x - solr supergroup 0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup 0 2022-03-17 06:48 /tmp
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:28 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /var
[cloudera@quickstart Desktop]$

```

Copy the file from one directory to other directory

```
[cloudera@quickstart:~]# hdfs dfs -cat /rjlocal/first01
This is a first file of Hadoop
[cloudera@quickstart:~]# hdfs dfs -mkdir /newdir
[cloudera@quickstart:~]# hdfs dfs -ls /
Found 8 items
drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup 0 2022-03-17 03:06 /hbase
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:53 /newdir
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:35 /rjlocal
drwxr-xr-x - solr solr 0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup 0 2022-03-17 06:48 /tmp
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:20 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /var
[cloudera@quickstart:~]# hdfs dfs -cp /rjlocal/first01 /newdir
[cloudera@quickstart:~]#
```

Check the newdir file is copy or not using "hdfs dfs -ls </dir_name>"

```
[cloudera@quickstart:~]# hdfs dfs -cat /rjlocal/first01
This is a first file of Hadoop
[cloudera@quickstart:~]# hdfs dfs -mkdir /newdir
[cloudera@quickstart:~]# hdfs dfs -ls /
Found 8 items
drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup 0 2022-03-17 03:06 /hbase
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:53 /newdir
drwxr-xr-x - solr solr 0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup 0 2022-03-17 06:48 /tmp
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:20 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /var
[cloudera@quickstart:~]# hdfs dfs -cp /rjlocal/first01 /newdir
[cloudera@quickstart:~]# hdfs dfs -ls /newdir
Found 1 items
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 07:58 /newdir/first01
[cloudera@quickstart:~]#
```

- 11. hdfs dfs -mv <dir1_name>/<file1_name> /output_abc moves file <file1_name> from source to destination within HDFS**

```
[cloudera@quickstart:~]# hdfs dfs -cat /rjlocal/first01
This is a first file of Hadoop
[cloudera@quickstart:~]# hdfs dfs -mv /rjlocal/first01 /output_01
[cloudera@quickstart:~]# hdfs dfs -ls /
Found 9 items
drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup 0 2022-03-17 03:06 /hbase
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:35 /newdir
drwxr-xr-x - solr solr 0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup 0 2022-03-17 06:48 /tmp
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:20 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /var
[cloudera@quickstart:~]# hdfs dfs -cp /rjlocal/first01 /newdir
[cloudera@quickstart:~]# hdfs dfs -ls /newdir
Found 1 items
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 07:58 /newdir/first01
[cloudera@quickstart:~]# hdfs dfs -ls /
Found 9 items
drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup 0 2022-03-17 03:06 /hbase
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:58 /newdir
drwxr-xr-x - solr solr 0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup 0 2022-03-17 06:40 /tmp
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:20 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /var
[cloudera@quickstart:~]#
```

After moving file source folder to destination folder

```
cloudera@quickstart:~/Desktop$ hdfs dfs -ls /newdir
Found 1 items
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 07:58 /newdir/first01
[cloudera@quickstart Desktop]$ hdfs dfs -mv /rjlocal/first01 /output_01
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 9 items
drwxr-xrwx - hdfs supergroup 0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup 0 2022-03-17 06:40 /tmp
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:20 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /var
[cloudera@quickstart Desktop]$ hdfs dfs -cp /rjlocal/first01 /newdir
[cloudera@quickstart Desktop]$ hdfs dfs -ls /newdir
Found 1 items
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 07:58 /newdir/first01
[cloudera@quickstart Desktop]$ hdfs dfs -mv /rjlocal/first01 /output_01
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 9 items
drwxr-xrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup 0 2022-03-17 03:06 /hbase
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:58 /newdir
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 07:35 /output_01
drwxr-xr-x - hdfs supergroup 0 2022-03-17 12:41 /rjlocal
drwxr-xr-x - solr supergroup 0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup 0 2022-03-17 06:40 /tmp
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:20 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /var
[cloudera@quickstart Desktop]$ hdfs dfs -ls /rjlocal
[cloudera@quickstart Desktop]$ hdfs dfs -ls /output_01
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 07:35 /output_01
[cloudera@quickstart Desktop]$
```

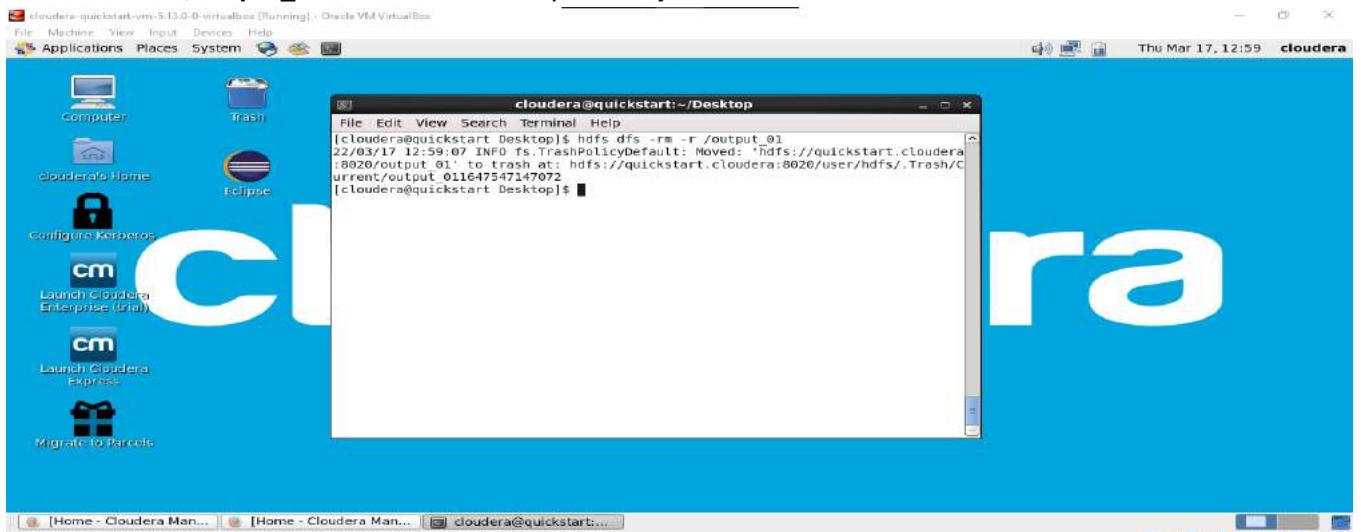
Use **hdfs dfs -cat </file1_name>** command to read moves file

```
cloudera@quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Thu Mar 17, 12:49 cloudera
cloudera@quickstart:~/Desktop$ hdfs dfs -cat /output_01
This is a first file of Hadoop.
[cloudera@quickstart Desktop]$
```

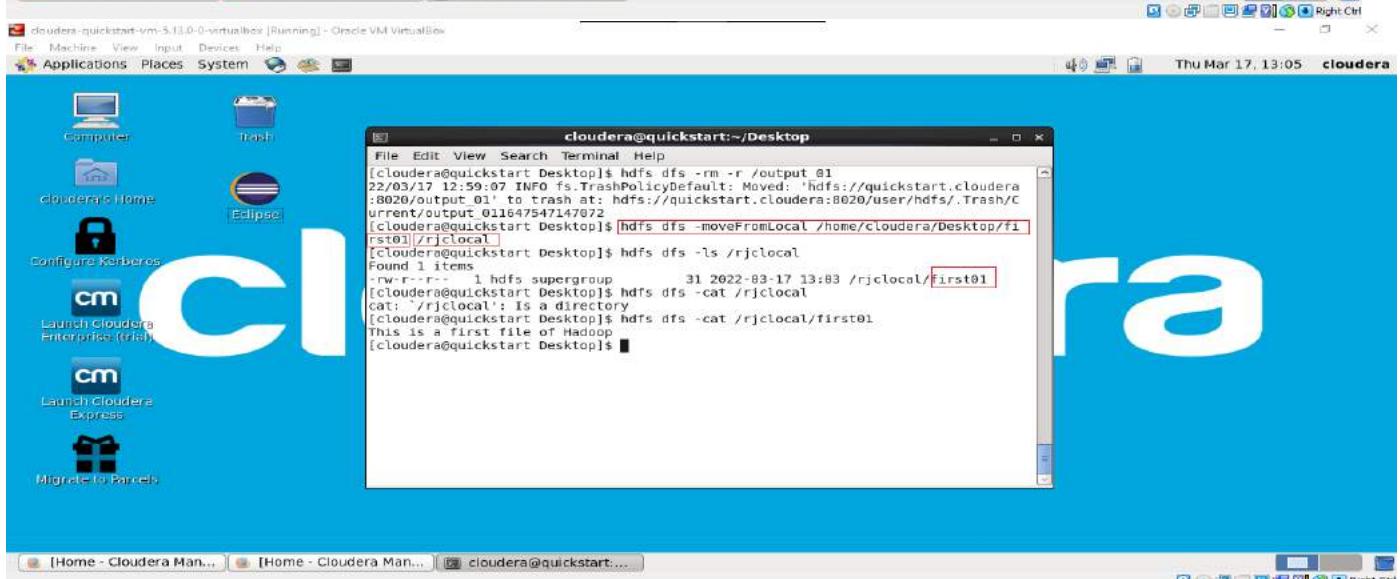
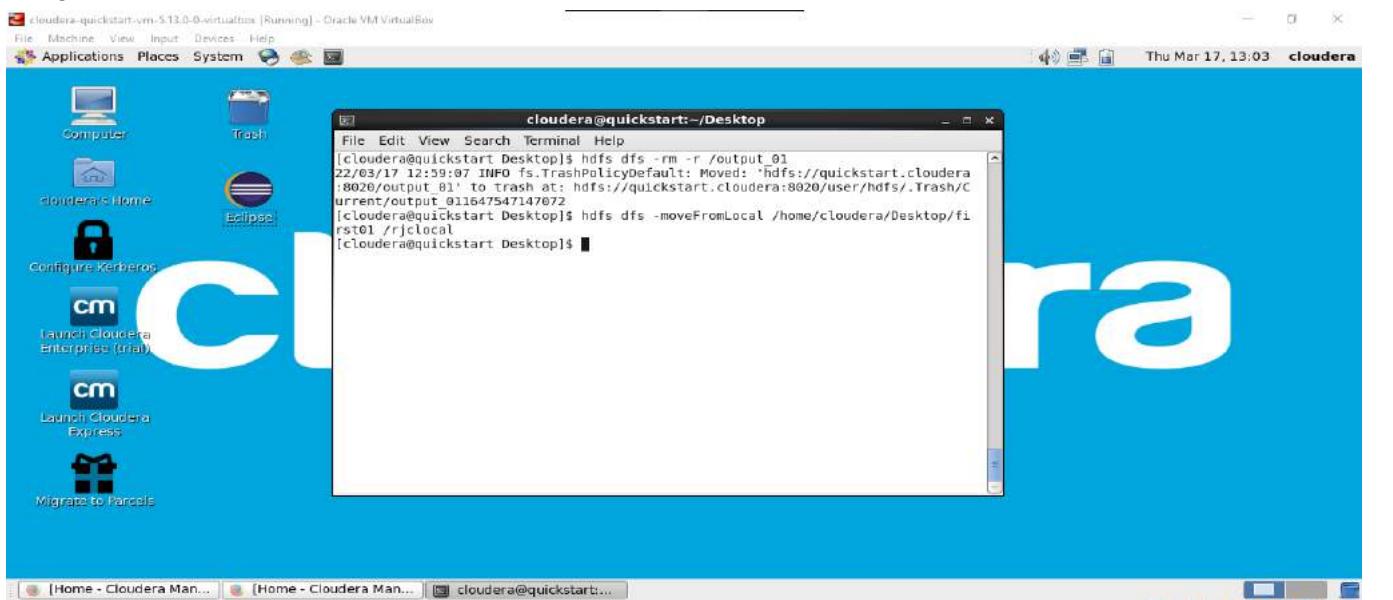
12. **hdfs dfs -rm </dir_name>/<file_name>** deletes objects and directories full of objects

```
cloudera@quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Thu Mar 17, 12:55 cloudera
cloudera@quickstart:~/Desktop$ hdfs dfs -rm /rjlocal/first01
[cloudera@quickstart Desktop]$ hdfs dfs -ls /newdir
Found 1 items
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 07:58 /newdir/first01
[cloudera@quickstart Desktop]$ hdfs dfs -ls /newdir
Found 9 items
drwxr-xrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup 0 2022-03-17 03:06 /hbase
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:58 /newdir
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 07:35 /output_01
drwxr-xr-x - hdfs supergroup 0 2022-03-17 12:41 /rjlocal
drwxr-xr-x - solr supergroup 0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup 0 2022-03-17 06:40 /tmp
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /var
[cloudera@quickstart Desktop]$ hdfs dfs -ls /rjlocal
[cloudera@quickstart Desktop]$ hdfs dfs -ls /output_01
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 07:35 /output_01
[cloudera@quickstart Desktop]$ hdfs dfs -cat /output_01
This is a first file of Hadoop.
[cloudera@quickstart Desktop]$ hdfs dfs -rm /rjlocal/first01
rm: '/rjlocal/first01': No such file or directory
[cloudera@quickstart Desktop]$
```

13. `hdfs dfs -rm -r /output_abc` removes directory with objects



14. `hdfs dfs -moveFromLocal /home/cloudera/Desktop/<file_name> /<dir_name>` moves files from local file system to HDFS



15. `hdfs dfs -moveToLocal <new_dir_name>/<file_name>` /home/cloudera/Desktop/<file_name> moves files.

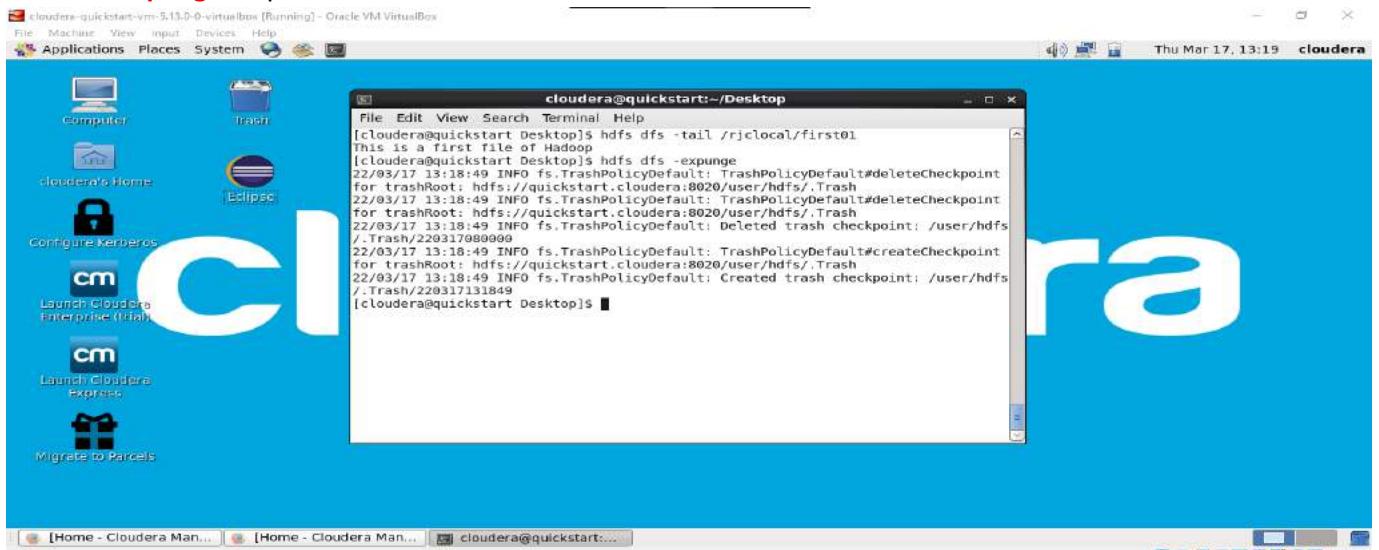
```
[cloudera@quickstart Desktop]$ hdfs dfs -rm -r /output_01
22/07/17 12:59:07 INFO fs.TrashPolicyDefault: Moved: 'hdfs://quickstart.cloudera:8020/output_01' to trash at: hdfs://quickstart.cloudera:8020/user/hdfs/.Trash/current/output_01
[cloudera@quickstart Desktop]$ hdfs dfs -moveFromLocal /home/cloudera/Desktop/first01 /rjlocal
[cloudera@quickstart Desktop]$ hdfs dfs -ls /rjlocal
Found 1 items
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 13:03 /rjlocal/first01
[cloudera@quickstart Desktop]$ hdfs dfs -cat /rjlocal/first01
cat: '/rjlocal': Is a directory
[cloudera@quickstart Desktop]$ hdfs dfs -cat /rjlocal/first01
This is a first file of Hadoop
[cloudera@quickstart Desktop]$ hdfs dfs -moveToLocal /new_dir/new_first01 /home/cloudera/Desktop/first01
moveToLocal: Option '-moveToLocal' is not implemented yet.
[cloudera@quickstart Desktop]$ hdfs dfs -moveToLocal /home/cloudera/Desktop/first01 /new_dir/new_first01
moveToLocal: Option '-moveToLocal' is not implemented yet.
[cloudera@quickstart Desktop]$
```

```
[cloudera@quickstart Desktop]$ hdfs dfs -cat /rjlocal
cat: '/rjlocal': Is a directory
[cloudera@quickstart Desktop]$ hdfs dfs -cat /rjlocal/first01
This is a first file of Hadoop
[cloudera@quickstart Desktop]$ hdfs dfs -moveToLocal /new_dir/new_first01 /home/cloudera/Desktop/first01
moveToLocal: Option '-moveToLocal' is not implemented yet.
[cloudera@quickstart Desktop]$ hdfs dfs -moveToLocal /home/cloudera/Desktop/first01 /new_dir/new_first01
moveToLocal: Option '-moveToLocal' is not implemented yet.
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 8 items
drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup 8 2022-03-17 03:06 /hbase
drwxr-xr-x - hdfs supergroup 8 2022-03-17 07:58 /newdir
drwxr-xr-x - hdfs supergroup 8 2022-03-17 13:03 /rjlocal
drwxr-xr-x - solr solr 0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup 8 2022-03-17 06:40 /tmp
drwxr-xr-x - hdfs supergroup 8 2022-03-17 07:20 /user
drwxr-xr-x - hdfs supergroup 8 2017-10-23 09:17 /var
[cloudera@quickstart Desktop]$ hdfs dfs -ls /newdir
Found 1 items
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 07:58 /newdir/first01
[cloudera@quickstart Desktop]$
```

16. `hdfs dfs -tail /<dir_name>/<file_name>` shows last 1KB of file

```
[cloudera@quickstart Desktop]$ hdfs dfs -tail /rjlocal/first01
This is a first file of Hadoop
[cloudera@quickstart Desktop]$
```

17. `hdfs dfs -expunge` empties the trash available in HDFS



18. First go to local host at `localhost:50070` click on Utilities and select browse the file system

Started:	Thu Mar 17 06:42:47 -0700 2022
Version:	2.6.0-cdh5.13.0, r42e8860b182e55321bd5f5605264da4adc8882be
Compiled:	Wed Oct 04 11:08:00 -0700 2017 by jenkins from Unknown
Cluster ID:	CID-a24185f9-a545-40fe-9553-84c3fdca489f
Block Pool ID:	BP-1067413441-127.0.0.1-1508775264580

Summary

Type / in test area and click go to see current directories in file system

Snapshot Summary

Snapshottable directories: 0

Path	Snapshot Number	Snapshot Quota	Modification Time	Permission	Owner	Group

Snapshotted directories: 0

Snapshot ID	Snapshot Directory	Modification Time

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxrwxrwx	hdfs	supergroup	0 B	Mon Oct 23 09:15:43 -0700 2017	0	0 B	benchmarks

hdfs dfs -setrep 4 </dir_name> sets replication of files in specifies directory 4 times

```
File Edit View Search Terminal Help
[cloudera@quickstart Desktop]$ hdfs dfs -setrep 4 /newdir
Replication 4 set: /newdir/first01
[cloudera@quickstart Desktop]$
```

Check replication of file within specified directory

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hdfs	supergroup	31 B	Thu Mar 17 07:58:45 -0700 2022	4	128 MB	first01

Browse Directory

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hdfs	supergroup	31 B	Thu Mar 17 07:58:45 -0700 2022	4	128 MB	first01

Hadoop, 2017.

19. `hdfs dfs -du /<dir_name>` gives size of a directory in HDFS

The screenshot shows a desktop environment with a terminal window and a file browser window. The terminal window displays the command `hdfs dfs -du /rjlocal` and its output, which shows a single directory named `first01` with a size of 128 MB. The file browser window shows a directory structure under `/newdir`, with one item named `first01`.

```
[cloudera@quickstart Desktop]$ hdfs dfs -ls /rjlocal
Found 1 items
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 13:03 /rjlocal/first01
[cloudera@quickstart Desktop]$ hdfs dfs -du /rjlocal
128 MB /rjlocal/first01
[cloudera@quickstart Desktop]$
```

Block Size	Name
128 MB	first01

20. `hdfs dfs -df` gives amount of space in use and space available on currently mounted file system

The screenshot shows a desktop environment with a terminal window displaying the command `hdfs dfs -df` and its output. The output shows a single filesystem entry for `hdfs://quickstart.cloudera:8020` with a total size of 58479091712, used space of 872663648, and available space of 45933694976, resulting in a 1% use percentage.

```
[cloudera@quickstart Desktop]$ hdfs dfs -df
Filesystem      Size        Used      Available  Use%
hdfs://quickstart.cloudera:8020  58479091712  872663648  45933694976   1%
[cloudera@quickstart Desktop]$
```

`hdfs dfs -df -h` gives space in gb and mb abbreviation

The screenshot shows a desktop environment with a terminal window displaying the command `hdfs dfs -df -h` and its output. The output shows a single filesystem entry for `hdfs://quickstart.cloudera:8020` with a total size of 54.5 G, used space of 832.2 M, and available space of 42.8 G, resulting in a 1% use percentage.

```
[cloudera@quickstart Desktop]$ hdfs dfs -df
Filesystem      Size        Used      Available  Use%
hdfs://quickstart.cloudera:8020  58479091712  872663648  45933694976   1%
[cloudera@quickstart Desktop]$ hdfs dfs -df -h
Filesystem      Size        Used      Available  Use%
hdfs://quickstart.cloudera:8020  54.5 G     832.2 M    42.8 G    1%
[cloudera@quickstart Desktop]$
```

- 21. hdfs fsck <dir_name> checks health of HDFS and moves corrupted files to lost+found directory or It delete a corrupted file present in HDFS**

```
[cloudera@quickstart Desktop]$ hdfs fsck /rjlocal
Connecting to namenode via http://quickstart.cloudera:50070/fsck?ugi=hdfs&path=%2Frjlocal
FSCK started by hdfs (auth:SIMPLE) from /10.0.2.15 for path /rjlocal at Fri Mar 18 05:29:41 PDT 2022
.Status: HEALTHY
Total size: 31 B
Total dirs: 1
Total files: 1
Total symlinks: 0
Total blocks (validated): 1 (avg. block size 31 B)
Minimally replicated blocks: 1 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Fri Mar 18 05:29:41 PDT 2022 in 14 milliseconds

The filesystem under path '/rjlocal' is HEALTHY
[cloudera@quickstart Desktop]$
```

hdfs fsck /rjlocal -files prints out the files being checked

```
[cloudera@quickstart Desktop]$ hdfs fsck /rjlocal -files
Connecting to namenode via http://quickstart.cloudera:50070/fsck?ugi=hdfs&path=%2Frjlocal
FSCK started by hdfs (auth:SIMPLE) from /10.0.2.15 for path /rjlocal at Fri Mar 18 05:32:45 PDT 2022
/rjlocal <dir>
/rjlocal/first01 31 bytes, 1 block(s): OK
.Status: HEALTHY
Total size: 31 B
Total dirs: 1
Total files: 1
Total symlinks: 0
Total blocks (validated): 1 (avg. block size 31 B)
Minimally replicated blocks: 1 (100.0 %)
Over-replicated blocks: 0 (0.0 %)
Under-replicated blocks: 0 (0.0 %)
Mis-replicated blocks: 0 (0.0 %)
Default replication factor: 1
Average block replication: 1.0
Corrupt blocks: 0
Missing replicas: 0 (0.0 %)
Number of data-nodes: 1
Number of racks: 1
FSCK ended at Fri Mar 18 05:32:45 PDT 2022 in 2 milliseconds

The filesystems under path '/rjlocal' is HEALTHY
[cloudera@quickstart Desktop]$
```

- 22. hdfs dfs -touchz <dir_name>/<empty_file_name> creates an empty file**

```
[cloudera@quickstart Desktop]$ hdfs dfs -ls /rjlocal
Found 1 items
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 13:03 /rjlocal/first01
[cloudera@quickstart Desktop]$ hdfs dfs -touchz /rjlocal/empFirst01
[cloudera@quickstart Desktop]$ hdfs dfs -ls /rjlocal
Found 2 items
-rw-r--r-- 1 hdfs supergroup 0 2022-03-18 05:36 /rjlocal/empFirst01
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 13:03 /rjlocal/first01
[cloudera@quickstart Desktop]$
```

23. `hdfs dfs -stat </dir_name>` prints statistics about the file or directory it shows recent date of modification

```
cloudera@quickstart:~$ hdfs dfs -stat /rjlocal
2022-03-18 12:36:11
```

`hdfs -stat %b </dir_name>/<file_name>` gives file size in bytes

```
cloudera@quickstart:~$ hdfs dfs -stat %b /rjlocal/first01
0
```

`hdfs dfs -stat %o /<dir_name>/<file_name>` gives block size of file

```
cloudera@quickstart:~$ hdfs dfs -stat %o /rjlocal/first01
134217728
```

`hdfs dfs -stat %r /<dir_name>/<file_name>` gives no. of replication of file

A screenshot of a Cloudera Quickstart VM desktop environment. The desktop has a blue background with large white text. On the left, there's a vertical bar with icons for Computer, Trash, cloudera Home, Eclipse, Configure Kerberos, Launch Cloudera Enterprise (trial), Launch Cloudera Express, and Migrate to Parcels. A large 'cloudera' logo is centered on the screen. At the top, the title bar shows 'cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox'. The menu bar includes File, Machine, View, Input, Devices, Help, Applications, Places, System, and Help. The system tray shows icons for battery, signal, and volume. The date and time 'Fri Mar 18, 05:51' are in the top right. A terminal window titled 'cloudera@quickstart:~/Desktop' is open, displaying the following command-line session:

```
[cloudera@quickstart Desktop]$ hdfs dfs -stat /rjlocal  
2022-03-18 12:36:11  
[cloudera@quickstart Desktop]$ hdfs dfs -ls /rjlocal  
Found 2 items:  
-rw-r--r-- 1 hdfs supergroup 0 2022-03-18 05:36 /rjlocal/empFirst01  
-rwxr--r-- 1 hdfs supergroup 31 2022-03-17 13:03 /rjlocal/first01  
[cloudera@quickstart Desktop]$ hdfs dfs -stat %/rjlocal/first01  
31  
[cloudera@quickstart Desktop]$ hdfs dfs -stat %/rjlocal/empFirst01  
0  
[cloudera@quickstart Desktop]$ hdfs dfs -stat %o /rjlocal/first01  
134217728  
[cloudera@quickstart Desktop]$ hdfs dfs -stat %o /rjlocal/empFirst01  
134217728  
[cloudera@quickstart Desktop]$ hdfs dfs -stat %r /rjlocal/first01  
1  
[cloudera@quickstart Desktop]$ hdfs dfs -stat %r /rjlocal/empFirst01  
1  
[cloudera@quickstart Desktop]$
```

hdfs dfs -stat %y /<dir_name>/<file_name> gives last modification date

The screenshot shows a desktop environment for Cloudera Quickstart VM. The desktop background features a large white 'cloudera' logo on a blue gradient background. On the left, there's a vertical dock with several icons: Computer, Trash, cloudera's Home, Configure Kerberos, Launch Cloudera Enterprise (trial), Launch Cloudera Express, and Migrate to Parcels. The main window is a terminal titled 'cloudera@quickstart:~/Desktop'. It displays the following command-line session:

```
cloudera@quickstart:~/Desktop
File Edit View Search Terminal Help
[cloudera@quickstart Desktop]$ hdfs dfs -stat /rjclocal
2022-03-18 12:36:11
[cloudera@quickstart Desktop]$ hdfs dfs -ls /rjclocal
Found 2 items
-rw-r--r-- 1 hdfs supergroup 0 2022-03-18 10:36 /rjclocal/empFirst01
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 13:03 /rjclocal/first01
[cloudera@quickstart Desktop]$ hdfs dfs -stat %b /rjclocal/first01
31
[cloudera@quickstart Desktop]$ hdfs dfs -stat %e /rjclocal/empFirst01
0
[cloudera@quickstart Desktop]$ hdfs dfs -stat %o /rjclocal/first01
134217728
[cloudera@quickstart Desktop]$ hdfs dfs -stat %o /rjclocal/empFirst01
134217728
[cloudera@quickstart Desktop]$ hdfs dfs -stat %r /rjclocal/first01
1
[cloudera@quickstart Desktop]$ hdfs dfs -stat %r /rjclocal/empFirst01
1
[cloudera@quickstart Desktop]$ hdfs dfs -stat %y /rjclocal/first01
2022-03-17 20:03:05
[cloudera@quickstart Desktop]$ hdfs dfs -stat %y /rjclocal/empFirst01
2022-03-18 12:36:11
[cloudera@quickstart Desktop]$
```

24. `hdfs dfs -checksum <dir_name>/<file_name>` gives checksum of file

25. hdfs dfs -help rm shows syntax of whereas commands

```
cloudera@quickstart:~$ hdfs dfs -help rm
[hdfs dfs -rm] <src> ...
Delete all files that match the specified file pattern. Equivalent to the Unix command "rm <src>".

-skipTrash option bypasses trash, if enabled, and immediately deletes <src>
-f If the file does not exist, do not display a diagnostic message or
modify the exit status to reflect an error.
-[rR] Recursively deletes directories
[cloudera@quickstart Desktop]$
```

26. hdfs dfs -get /<dir_name>/<file_name> /home/cloudera/Desktop copies file from HDFS to local file system

```
[cloudera@quickstart Desktop]$ hdfs dfs -get /rjlocal/file01 /home/cloudera/Desktop
get: '/rjlocal/file01': No such file or directory
[cloudera@quickstart Desktop]$ hdfs dfs -ls /rjlocal
Found 2 items
-rw-r--r-- 1 hdfs supergroup 0 2022-03-18 05:36 /rjlocal/empFirst01
-rw-r--r-- 1 hdfs supergroup 31 2022-03-17 13:03 /rjlocal/first01
[cloudera@quickstart Desktop]$ hdfs dfs -get /rjlocal/first01 /home/cloudera/Desktop
[cloudera@quickstart Desktop]$
```

Aim: To Implement WordCount problem using Hadoop MapReduce in Eclipse: (Without Combiner & With Combiner)

What is MapReduce?

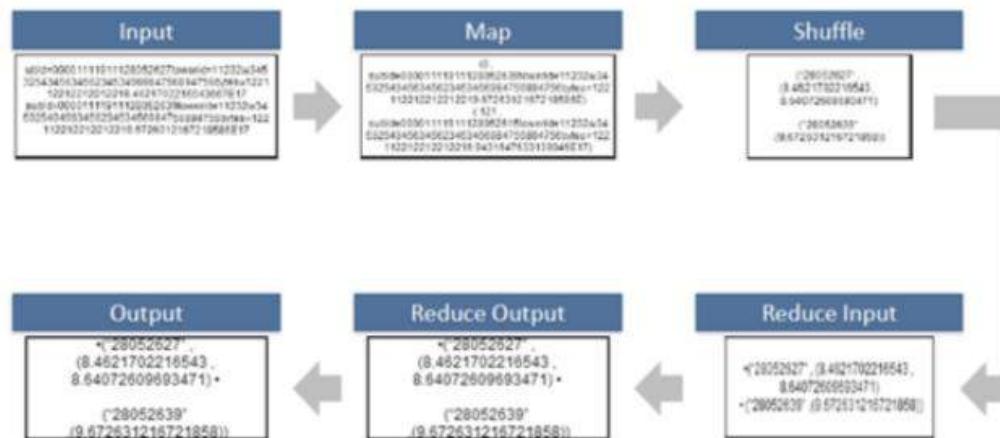
A MapReduce is a data processing tool which is used to process the data parallelly in a distributed form. It was developed in 2004, on the basis of paper titled as "MapReduce:

"Simplified Data Processing on Large Clusters," published by Google.

The MapReduce is a paradigm which has two phases, the mapper phase, and the reducer phase. In the Mapper, the input is given in the form of a key-value pair. The output of the Mapper is fed to the reducer as input. The reducer runs only after the Mapper is over. The reducer too takes input in key-value format, and the output of reducer is the final output.

Steps in Map Reduce

- The map takes data in the form of pairs and returns a list of <key, value> pairs. The keys will not be unique in this case.
- Using the output of Map, sort and shuffle are applied by the Hadoop architecture. This sort and shuffle acts on these list of <key, value> pairs and sends out unique keys and a list of values associated with this unique key <key, list(values)>.
- An output of sort and shuffle sent to the reducer phase. The reducer performs a defined function on a list of values for unique keys, and Final output <key, value> will be stored/displayed.

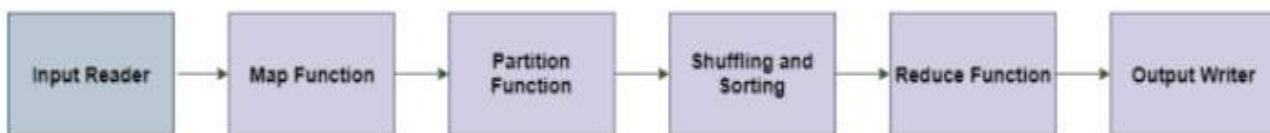


Sort and Shuffle

The sort and shuffle occur on the output of Mapper and before the reducer. When the Mapper task is complete, the results are sorted by key, partitioned if there are multiple reducers, and then written to disk. Using the input from each Mapper <k2,v2>, we collect all the values for each unique key k2. This output from the shuffle phase in the form of <k2, list(v2)> is sent as input to reducer phase.

Data Flow In MapReduce

MapReduce is used to compute the huge amount of data. To handle the upcoming data in a parallel and distributed form, the data has to flow from various phases.



Phases of MapReduce data flow

Input reader

The input reader reads the upcoming data and splits it into the data blocks of the appropriate size (64 MB to 128 MB). Each data block is associated with a Map function.

Once input reads the data, it generates the corresponding key-value pairs. The input files reside in HDFS.

Map function

The map function process the upcoming key-value pairs and generated the corresponding output key-value pairs. The map input and output type may be different from each other.

Partition function

The partition function assigns the output of each Map function to the appropriate reducer. The available key and value provide this function. It returns the index of reducers.

Shuffling and Sorting

The data are shuffled between/within nodes so that it moves out from the map and get ready to process for reduce function. Sometimes, the shuffling of data can take much computation time.

The sorting operation is performed on input data for Reduce function. Here, the data is compared using comparison function and arranged in a sorted form.

Reduce function

The Reduce function is assigned to each unique key. These keys are already arranged in sorted order. The values associated with the keys can iterate the Reduce and generates the corresponding output.

Output writer

Once the data flow from all the above phases, Output writer executes. The role of Output writer is to write the Reduce output to the stable storage.

To Implement WordCount problem using Hadoop MapReduce in Eclipse:

Hadoop WordCount operation occurs in 3 stages –

- Mapper Phase
- Shuffle Phase
- Reducer Phase

Hadoop WordCount - Mapper Phase Execution

- The text from the input text file is tokenized into words to form a key value pair with all the words present in the input text file. The key is the word from the input file and value is ‘1’.
- For instance if you consider the sentence “An elephant is an animal”.
- The mapper phase in the WordCount example will split the string into individual tokens i.e. words. In this case, the entire sentence will be split into 5 tokens (one for each word) with a value 1 as shown below –

Key-Value pairs from Hadoop Map Phase Execution-

(an,1)
(elephant,1)
(is,1)
(an,1)
(animal,1)

Hadoop WordCount Example- Shuffle Phase Execution

- After the map phase execution is completed successfully, shuffle phase is executed automatically wherein the key-value pairs generated in the map phase are taken as input and then sorted in alphabetical order.

- After the shuffle phase is executed from the WordCount example code, the output will look like this –

(an,1)
(an,1)
(animal,1)
(elephant,1)
(is,1)

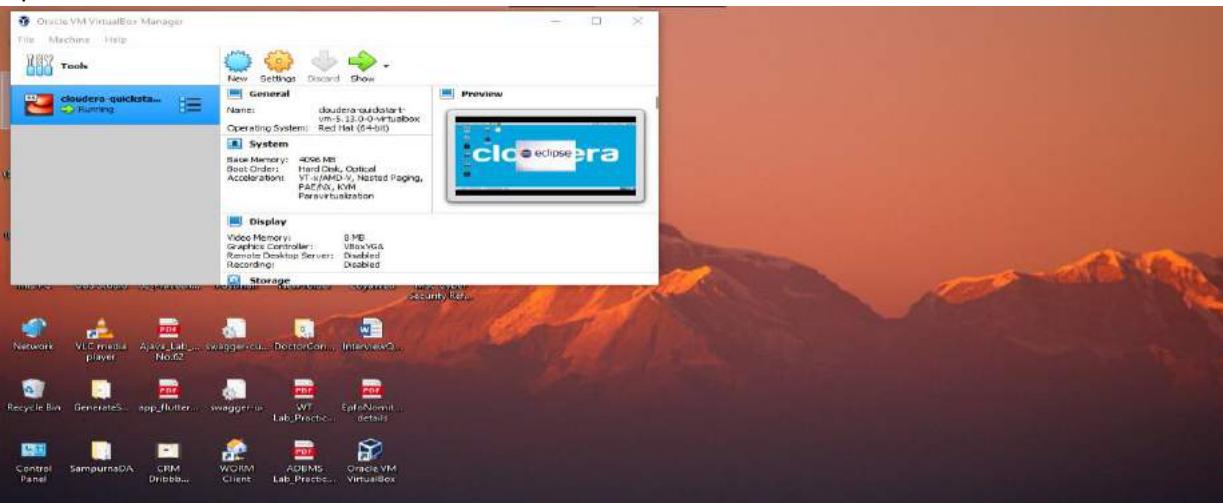
Hadoop WordCount Example- Reducer Phase Execution

- In the reduce phase, all the keys are grouped together and the values for similar keys are added up to find the occurrences for a particular word.

- It is like an aggregation phase for the keys generated by the map phase. The reducer phase takes the output of shuffle phase as input and then reduces the key-value pairs to unique keys with values added up.
- In our example “An elephant is an animal.” is the only word that appears twice in the sentence.
- After the execution of the reduce phase of MapReduce WordCount example program, appears as a key only once but with a count of 2 as shown below –
(an,2)
(animal,1)
(elephant,1)
(is,1)
- This is how the MapReduce word count program executes and outputs the number of occurrences of a word in any given input file.
- An important point to note during the execution of the WordCount example is that the mapper class in the WordCount program will execute completely on the entire input file and not just a single sentence.
- Suppose if the input file has 15 lines then the mapper class will split the words of all the 15 lines and form initial key value pairs for the entire dataset.
- The reducer execution will begin only after the mapper phase is executed successfully.

Steps for Word Count in Cloudera: (Without Combiner)

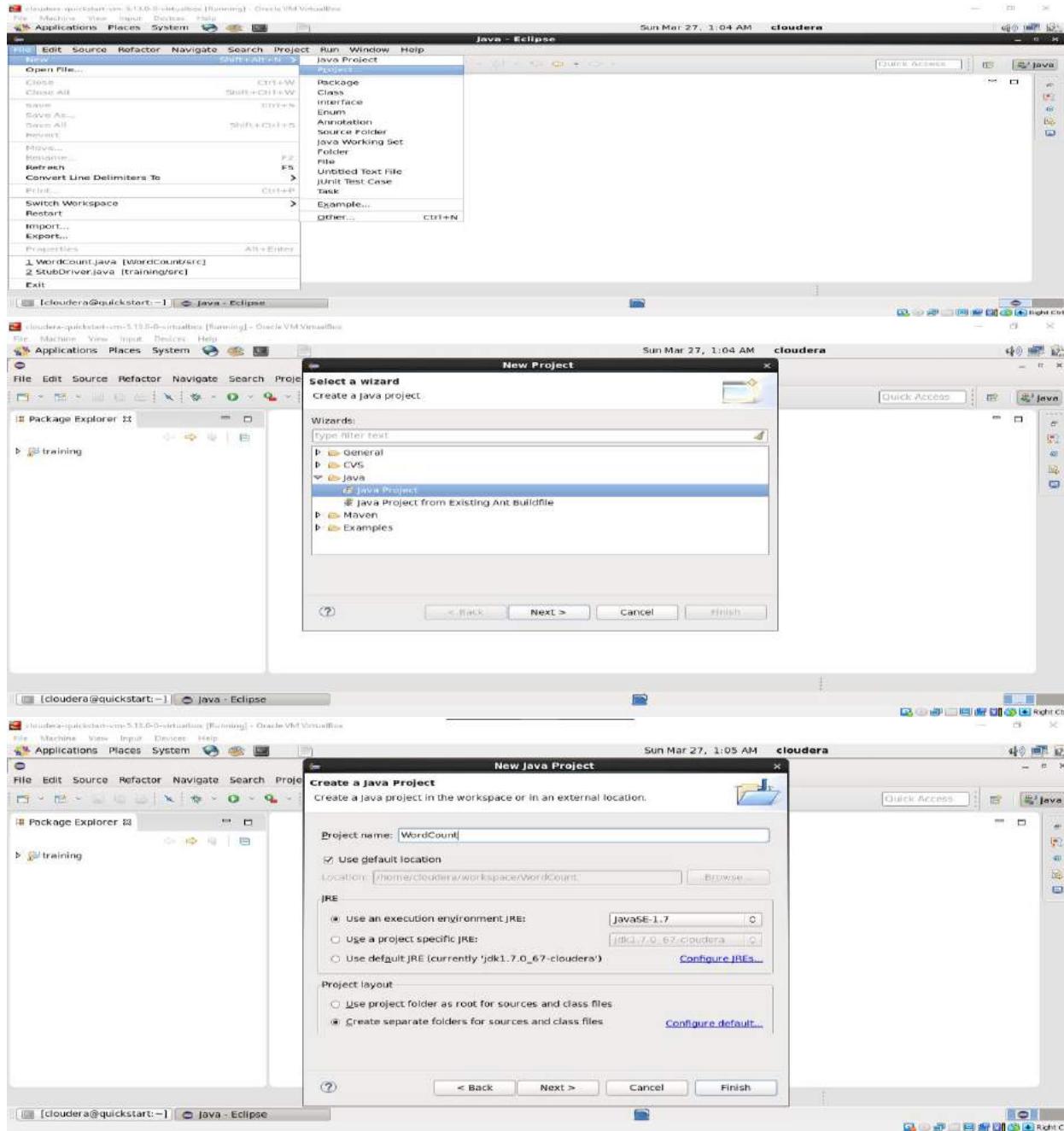
1. Open VirtualBox and then start Cloudera VM



2. Open Eclipse

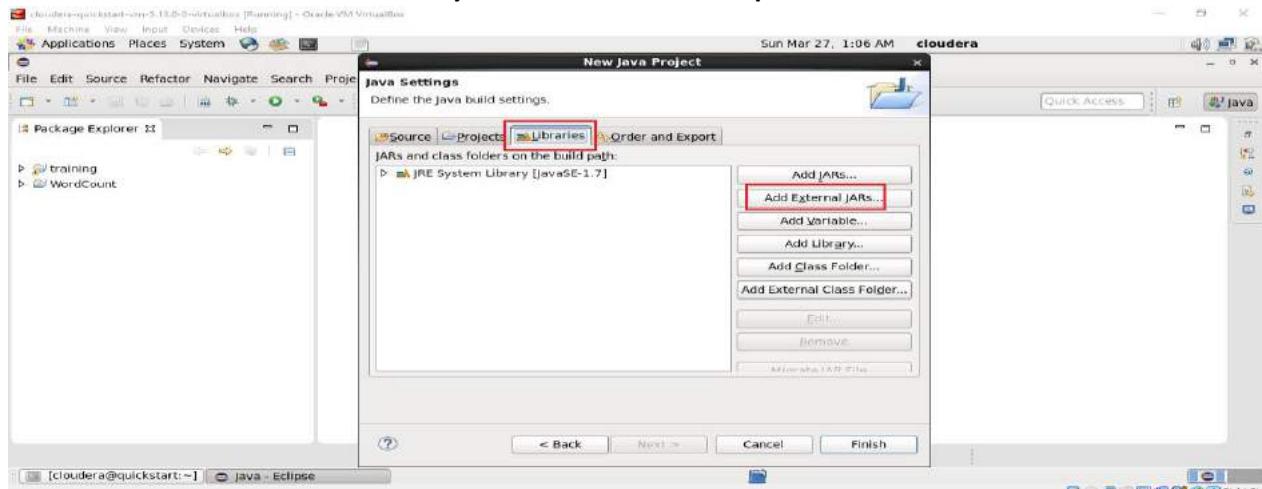


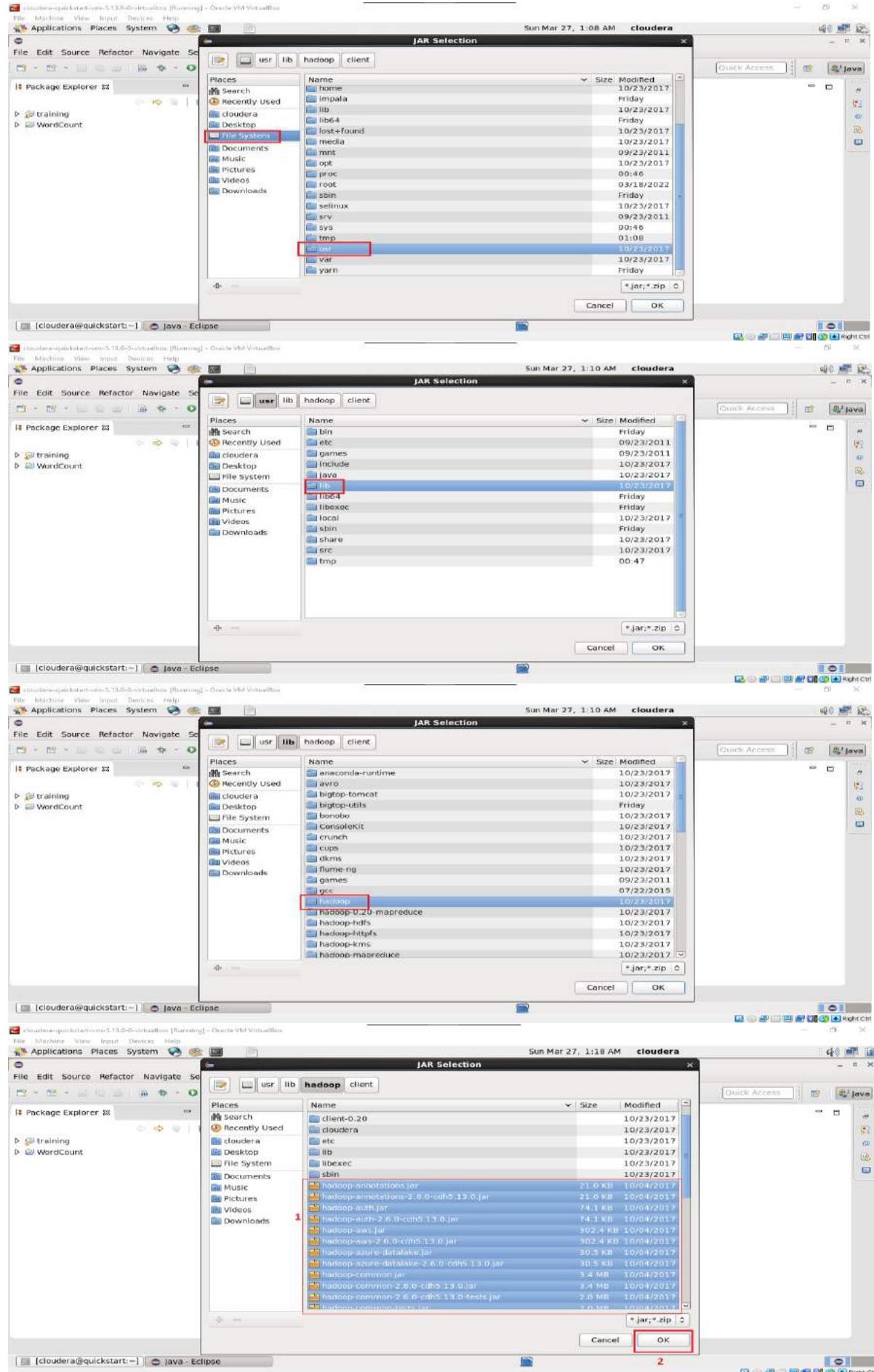
3. Create a New Java Project: File > New > Project > Java Project > Next Set project name as "WordCount".

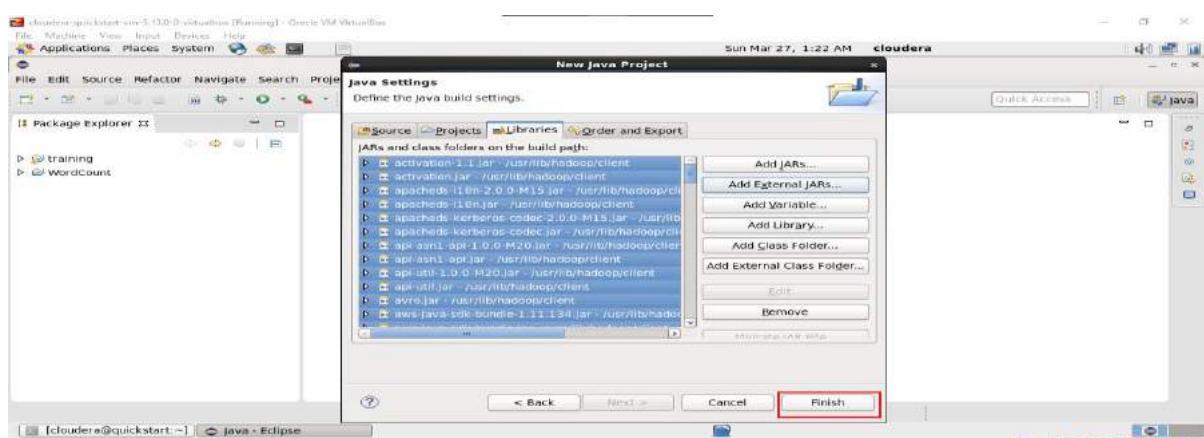
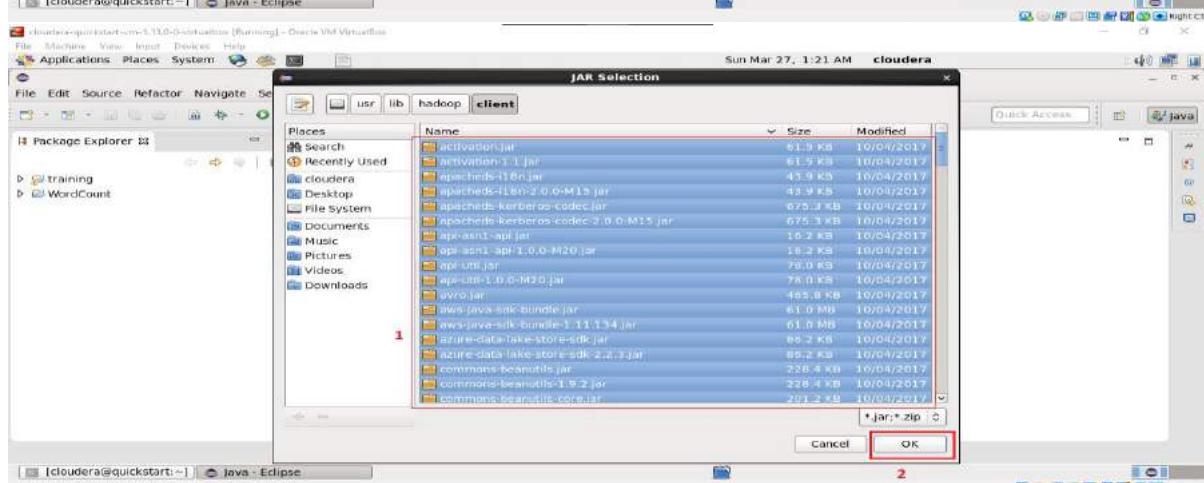
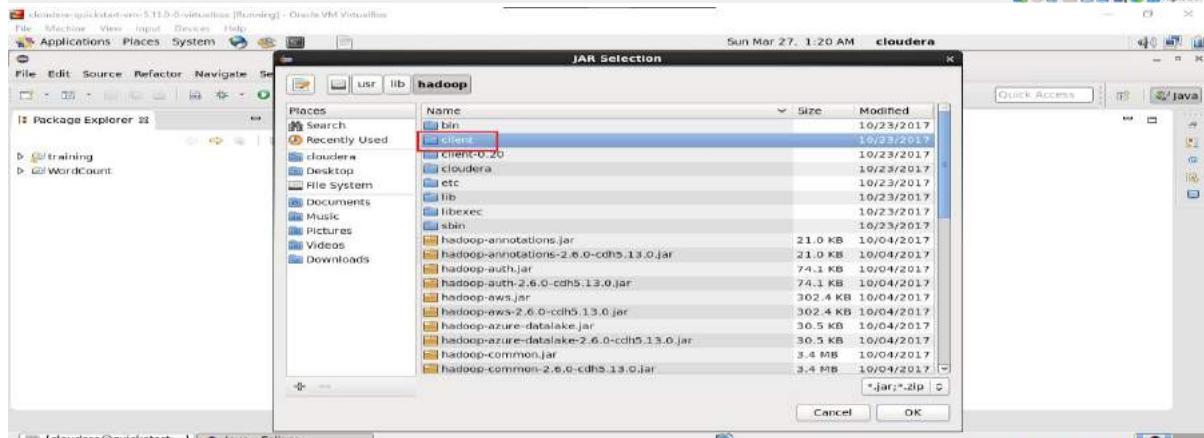
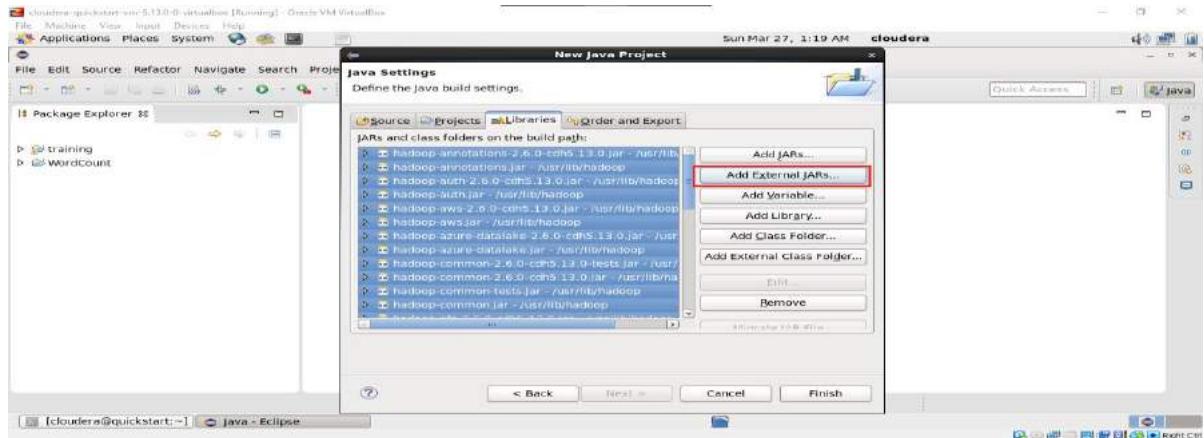


4. Adding the hadoop libraries to the project click on:

Libraries > Add External JARs > File System > usr > lib > hadoop > client > select all JAR files > OK > Finish.

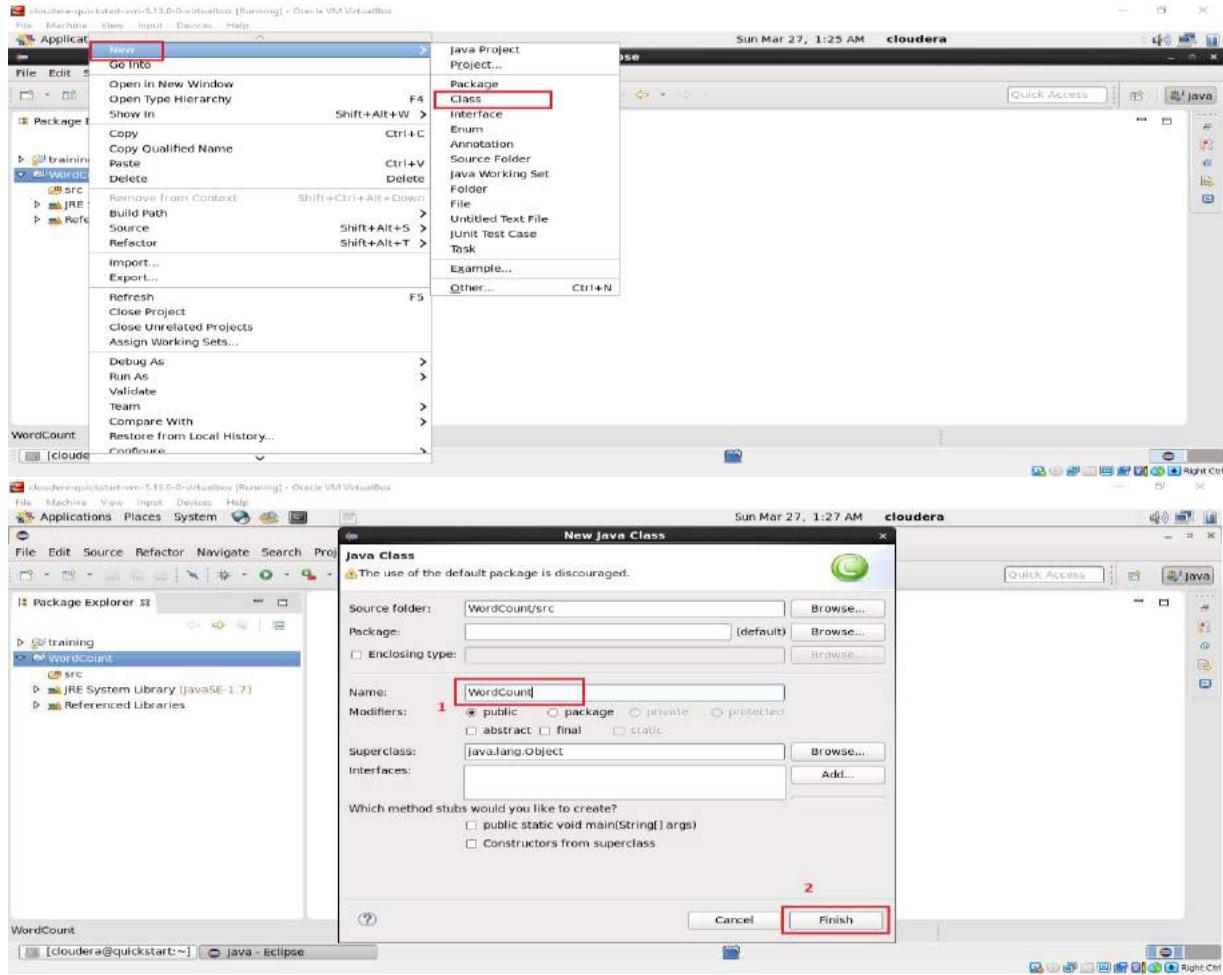






5. Right click on the name of project

"WordCount" > New > Class (Don't write anything for package) > class name "WordCount" > Finish
Then WordCount.java will open if not then open WordCount.java file manually



6. Source Code:

```

1 //Packages
2+ import java.io.IOException;[]
16
17 public class WordCount {
18     //Map Logic
19+     public static class TokenizerMapper extends Mapper<Object, Text, IntWritable> {}[]
31
32     //Reducer
33+     public static class IntSumReducer extends Reducer<Text, IntWritable, Text, IntWritable> {}[]
46
47     //Main Function
48+     public static void main(String[] args) throws Exception {}[]
62 }
63

```

```

//Packages

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

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

public class WordCount {

```

```
//Map Logic
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);
        }
    }
}

//Reducer
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);
    }
}

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

```

1 //Packages
2 import java.io.IOException;
3 import java.util.StringTokenizer;
4
5 import org.apache.hadoop.conf.Configuration;
6 import org.apache.hadoop.fs.Path;
7 import org.apache.hadoop.io.IntWritable;
8 import org.apache.hadoop.io.Text;
9
10 import org.apache.hadoop.mapreduce.Job;
11 import org.apache.hadoop.mapreduce.Mapper;
12 import org.apache.hadoop.mapreduce.Reducer;
13
14 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
15 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
16

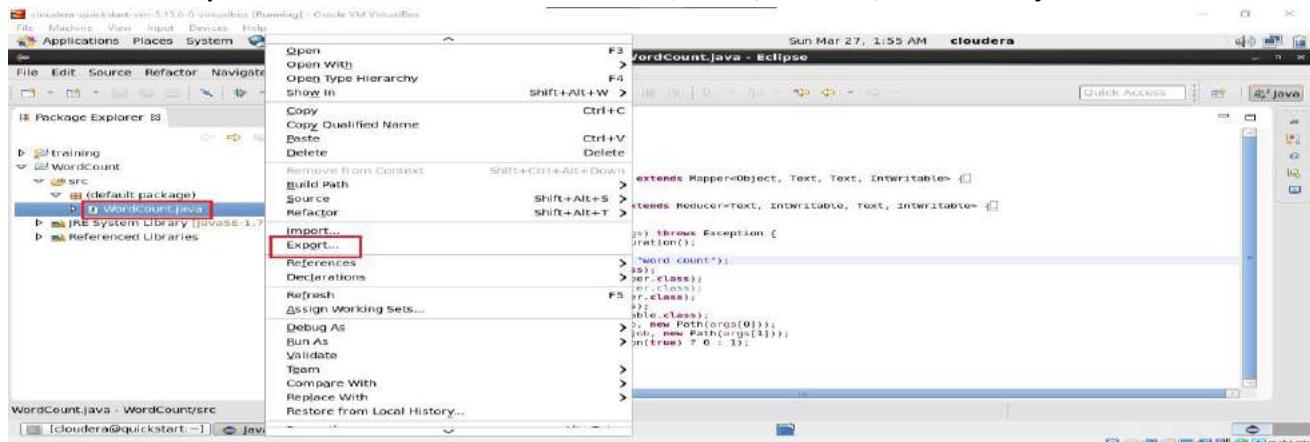
18 //Map Logic
19 public static class TokenizerMapper extends Mapper<Object, Text, Text, IntWritable> {
20     private final static IntWritable one = new IntWritable(1);
21     private Text word = new Text();
22
23     public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
24         StringTokenizer itr = new StringTokenizer(value.toString());
25         while (itr.hasMoreTokens()) {
26             word.set(itr.nextToken());
27             context.write(word, one);
28         }
29     }
30 }
31
32 //Reducer
33 public static class IntSumReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
34     private IntWritable result = new IntWritable();
35
36     public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException {
37         int sum = 0;
38         for (IntWritable val : values) {
39             sum += val.get();
40         }
41         result.set(sum);
42         context.write(key, result);
43     }
44 }
45
46
47 //Main Function
48 public static void main(String[] args) throws Exception {
49     Configuration conf = new Configuration();
50
51     Job job = Job.getInstance(conf, "word count");
52     job.setJarByClass(WordCount.class);
53     job.setMapperClass(TokenizerMapper.class);
54     //job.setCombinerClass(IntSumReducer.class);
55     job.setReducerClass(IntSumReducer.class);
56     job.setOutputKeyClass(Text.class);
57     job.setOutputValueClass(IntWritable.class);
58     FileInputFormat.addInputPath(job, new Path(args[0]));
59     FileOutputFormat.setOutputPath(job, new Path(args[1]));
60     System.exit(job.waitForCompletion(true) ? 0 : 1);
61 }

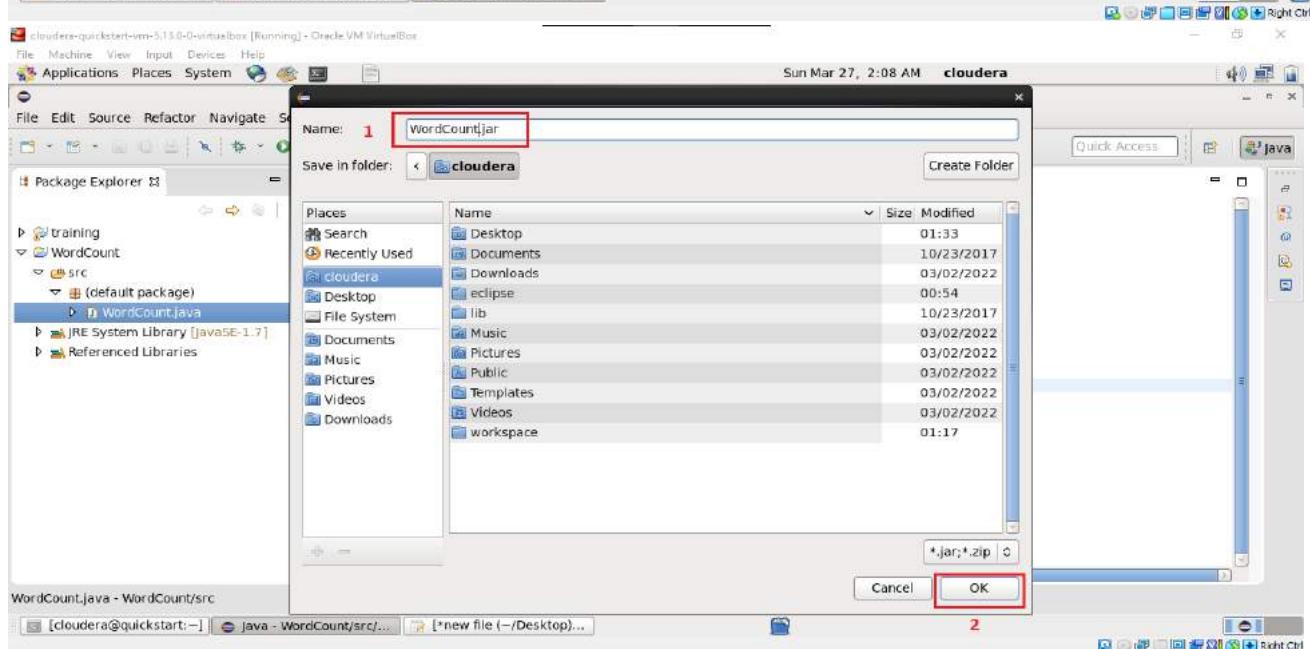
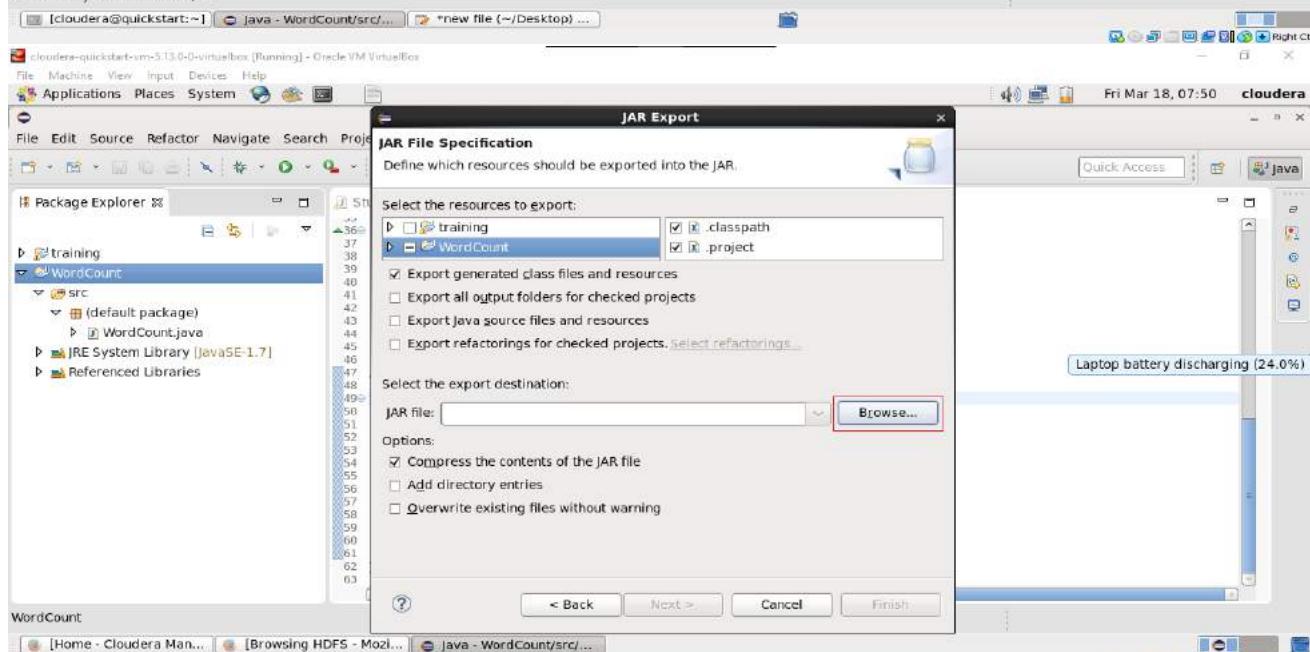
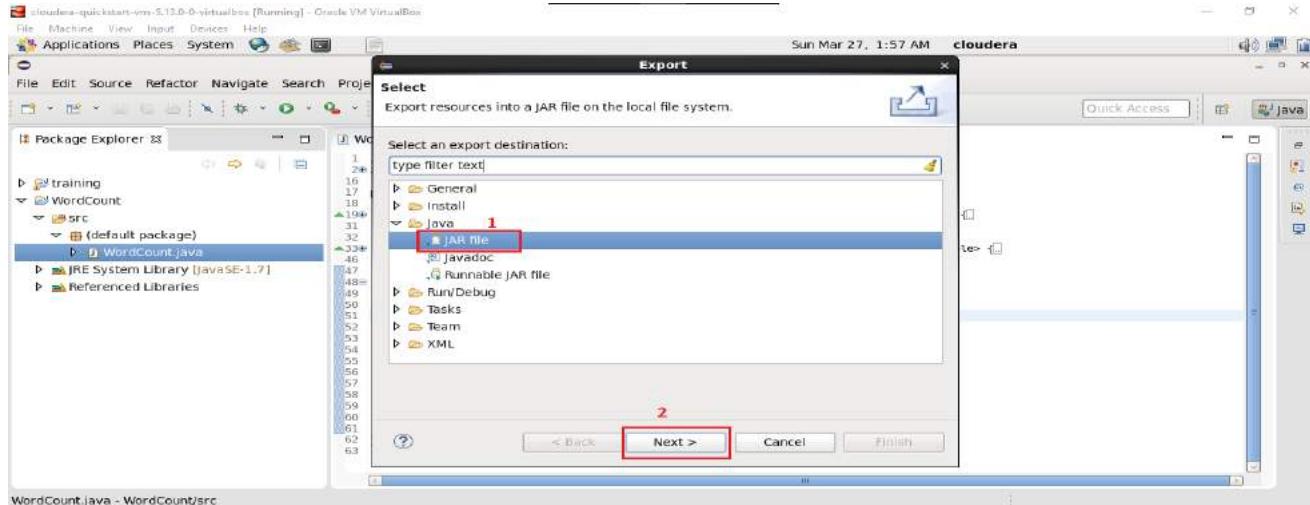
```

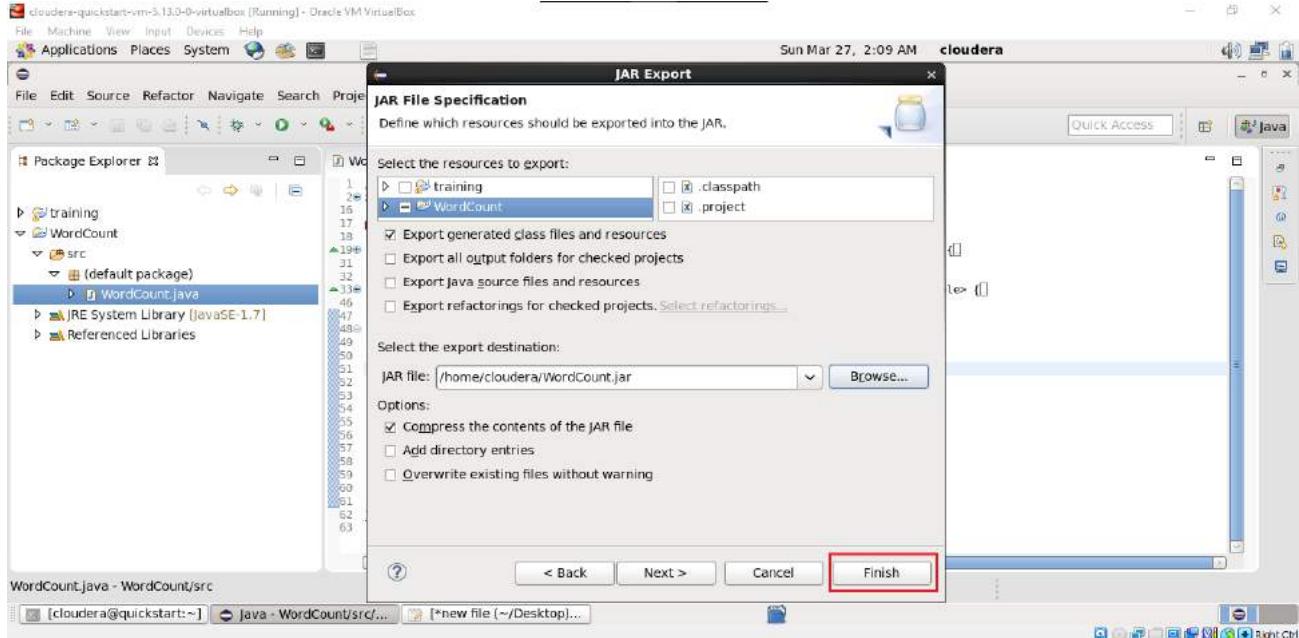
Note: We are running the code without combiner. That is why we commented the combiner line in main function.

7. Right click on the project name

WordCount > Export > Java > JAR File > Next > Jar file “/home/cloudera/WordCount.jar” > Finish.







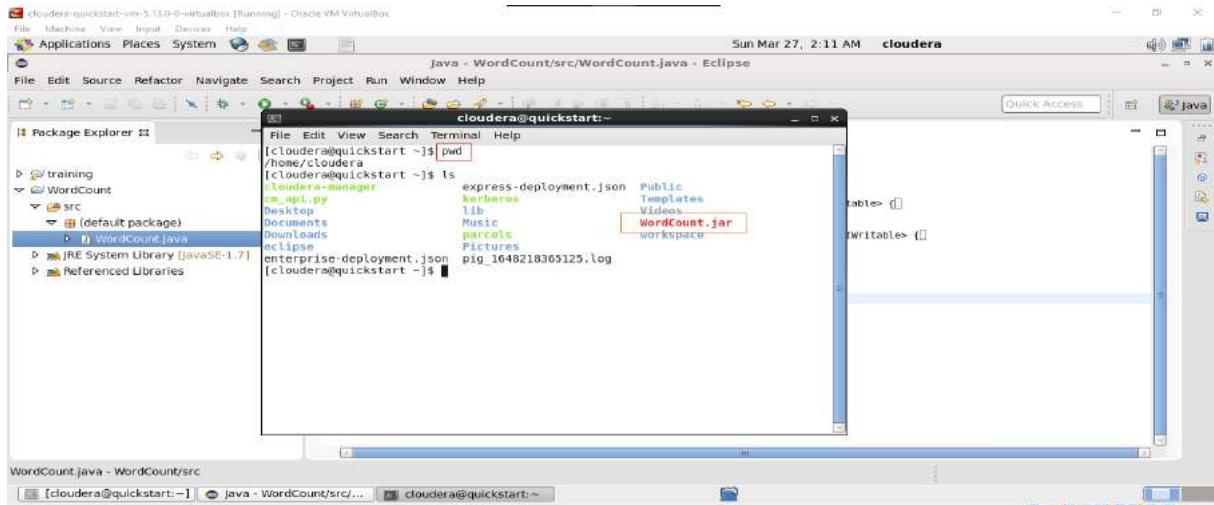
Verify jar file from terminal by using Open terminal & type "ls" There it will show WordCount.jar

Check current working directory

pwd

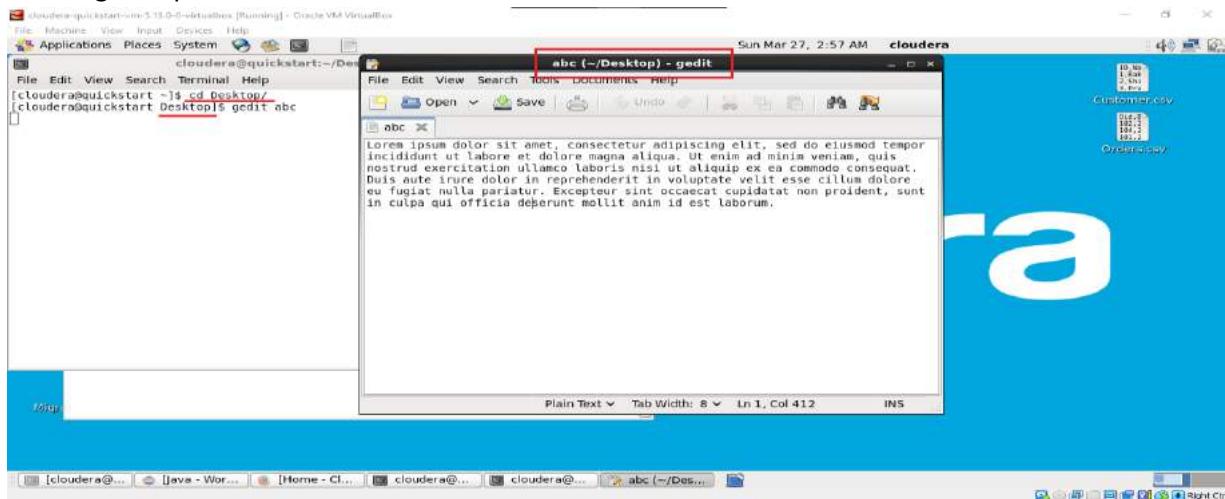
Check the list

ls



8. We need to create an input file in local file system (On Desktop)

Creating an input file named as "abc".



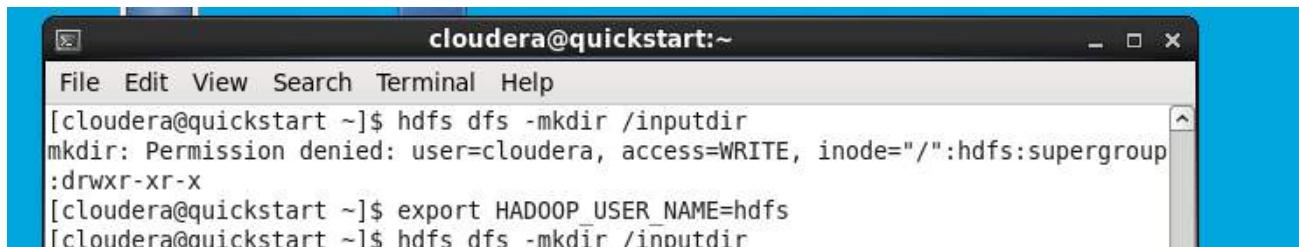
Here listing all the directory present in

hdfs using hdfs dfs -ls /

```
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 8 items
drwxrwxrwx  - hdfs  supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x  - hbase supergroup          0 2022-03-27 00:51 /hbase
drwxr-xr-x  - hdfs  supergroup          0 2022-03-17 07:58 /newdir
drwxr-xr-x  - hdfs  supergroup          0 2022-03-18 05:36 /rjclokal
drwxr-xr-x  - solr   solr              0 2017-10-23 09:18 /solr
drwxrwxrwt  - hdfs  supergroup          0 2022-03-25 08:12 /tmp
drwxr-xr-x  - hdfs  supergroup          0 2022-03-17 07:20 /user
drwxr-xr-x  - hdfs  supergroup          0 2017-10-23 09:17 /var
[cloudera@quickstart ~]$
```

9. Now we have to move this input file to hdfs. For this we create a direcory on hdfs using command

hdfs dfs -mkdir /inputdir



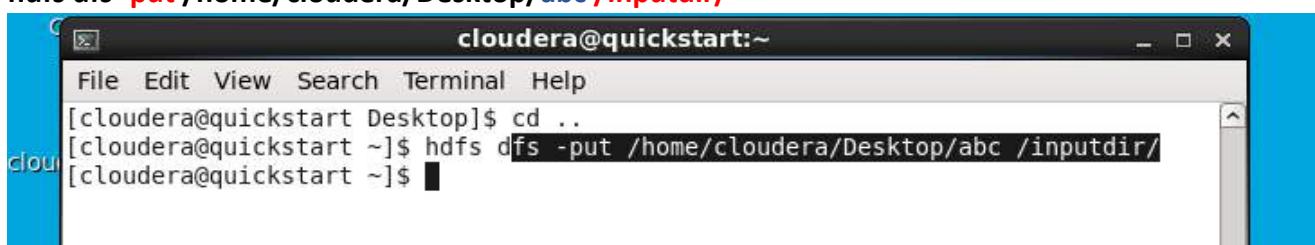
Then we can verify whether this directory is created or not using ls command

hdfs dfs -ls /

```
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 9 items
drwxrwxrwx  - hdfs  supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x  - hbase supergroup          0 2022-03-27 02:30 /hbase
drwxr-xr-x  - hdfs  supergroup          0 2022-03-27 02:46 /inputdir
drwxr-xr-x  - hdfs  supergroup          0 2022-03-17 07:58 /newdir
drwxr-xr-x  - hdfs  supergroup          0 2022-03-18 05:36 /rjclokal
drwxr-xr-x  - solr   solr              0 2017-10-23 09:18 /solr
drwxrwxrwt  - hdfs  supergroup          0 2022-03-25 08:12 /tmp
drwxr-xr-x  - hdfs  supergroup          0 2022-03-17 07:20 /user
drwxr-xr-x  - hdfs  supergroup          0 2017-10-23 09:17 /var
[cloudera@quickstart ~]$
```

Move the input file to this directory created in hdfs by using either **put** command or **copyFromLocal** command.

hdfs dfs -put /home/cloudera/Desktop/abc /inputdir/



Now checking whether the “abc” present in /inputdir directory of hdfs or not using command

hdfs dfs –ls /inputdir

```
[cloudera@quickstart ~]$ hdfs dfs -ls /inputdir/
Found 1 items
-rw-r--r-- 1 hdfs supergroup      446 2022-03-27 02:55 /inputdir/abc
[cloudera@quickstart ~]$
```

As we can see “abc” file is present in /**inputdir** directory of hdfs. Now we will see the content of this file using command

hdfs dfs –cat /inputdir/abc

```
[cloudera@quickstart ~]$ hdfs dfs -cat /inputdir/abc
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor i-
ncididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostru-
d exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aut-
e irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat n-
ulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui-
officia deserunt mollit anim id est laborum.
[cloudera@quickstart ~]$
```

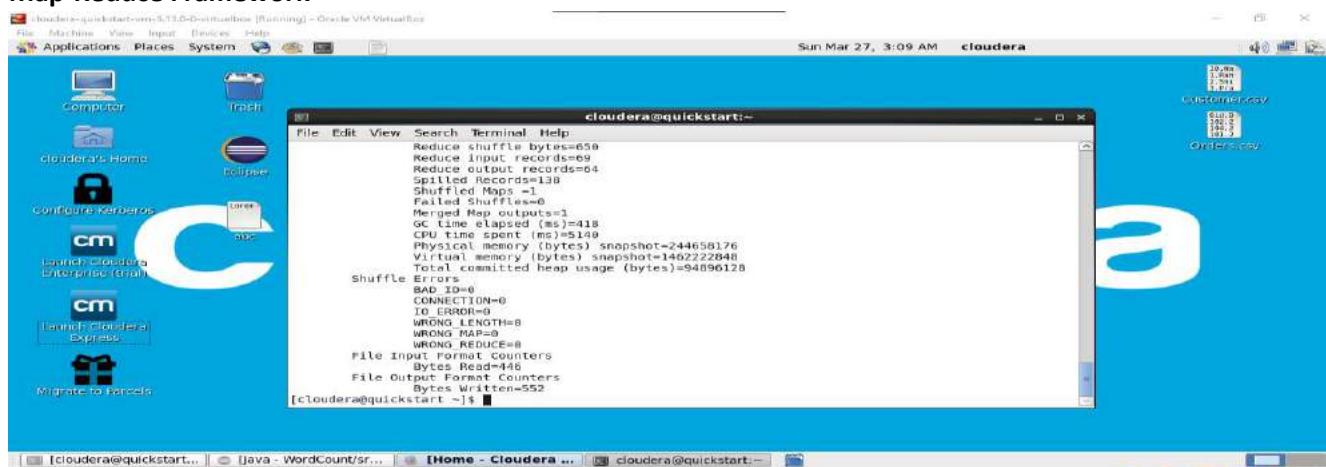
10. Running Mapreduce Program on Hadoop,

Syntax : **hadoop jar jarFileName.jar ClassName /InputFileAddress /outputdir**

```
hadoop jar /home/cloudera/WordCount.jar WordCount /inputdir/abc /outputdir
```

```
[cloudera@quickstart ~]$ hadoop jar /home/cloudera/WordCount.jar WordCount /inputdir/abc /outputdir  
22/03/27 03:05:46 INFO client.RMProxy: Connecting to ResourceManager at quickstart.cloudera/10.0.2.15:  
8032
```

Map-Reduce Framework



As we can see in the above output,

Combine input records=0

Combine output records=0

Note: We are getting this because we have **commented the Combiner line in main function.**

And Reduce shuffle bytes coming as,

Reduce shuffle bytes=1876

So when we are not using combiner 1876 bytes acting as an input for the reducer.

11. Then we can verify the content of tempop1 directory and in that part-r file has the actual output by using the command Hdfs dfs -cat /tempop1/part-r-00000 This will give us final output. The same file can also be accessed using a browser. For every execution of this program we need to delete the output directory or give a new name to the output directory every time. 1st we are checking whether the tempop1 directory is created in hdfs or not using command

hdfs dfs -ls /

```
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 10 items
drwxrwxrwx  -  hdfs  supergroup  0 2017-10-23 09:15 /benchmarks
drwxr-xr-x  -  hbase supergroup  0 2022-03-27 02:30 /hbase
drwxr-xr-x  -  hdfs  supergroup  0 2022-03-27 02:55 /inputdir
drwxr-xr-x  -  hdfs  supergroup  0 2022-03-17 07:58 /newdir
drwxr-xr-x  -  hdfs  supergroup  0 2022-03-27 03:08 /outputdir
drwxr-xr-x  -  hdfs  supergroup  0 2022-03-18 05:36 /rjclocal
drwxr-xr-x  -  solr   solr      0 2017-10-23 09:18 /solr
drwxrwxrwt  -  hdfs  supergroup  0 2022-03-25 08:12 /tmp
drwxr-xr-x  -  hdfs  supergroup  0 2022-03-17 07:20 /user
drwxr-xr-x  -  hdfs  supergroup  0 2017-10-23 09:17 /var
[cloudera@quickstart ~]$
```

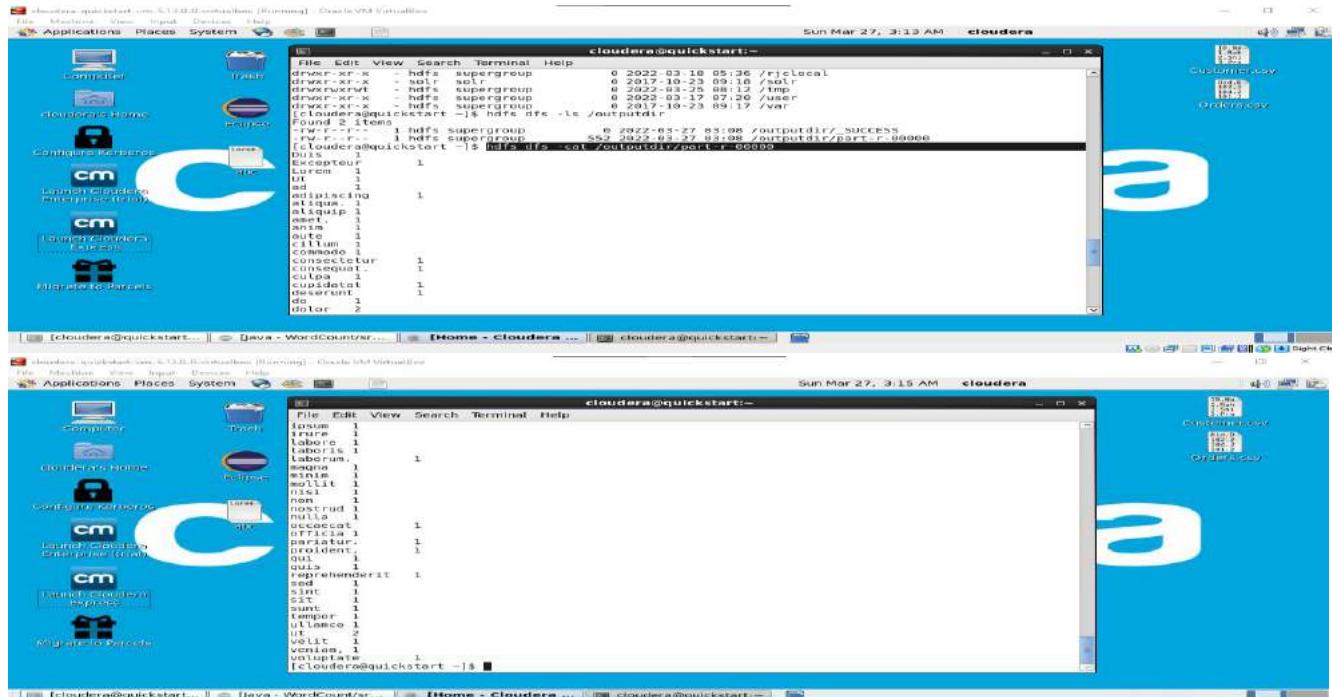
Now let's check what we have inside this **outputdir** directory using command as

hdfs dfs –ls /outputdir

```
[cloudera@quickstart ~]$ hdfs dfs -ls /outputdir
Found 2 items
-rw-r--r--  1 hdfs supergroup          0 2022-03-27 03:08 /outputdir/_SUCCESS
-rw-r--r--  1 hdfs supergroup      552 2022-03-27 03:08 /outputdir/part-r-00000
[cloudera@quickstart ~]$
```

Now we want to read the content of the part-r-00000 file which present inside the outputdir using command

hdfs dfs -cat /outputdir/part-r-00000

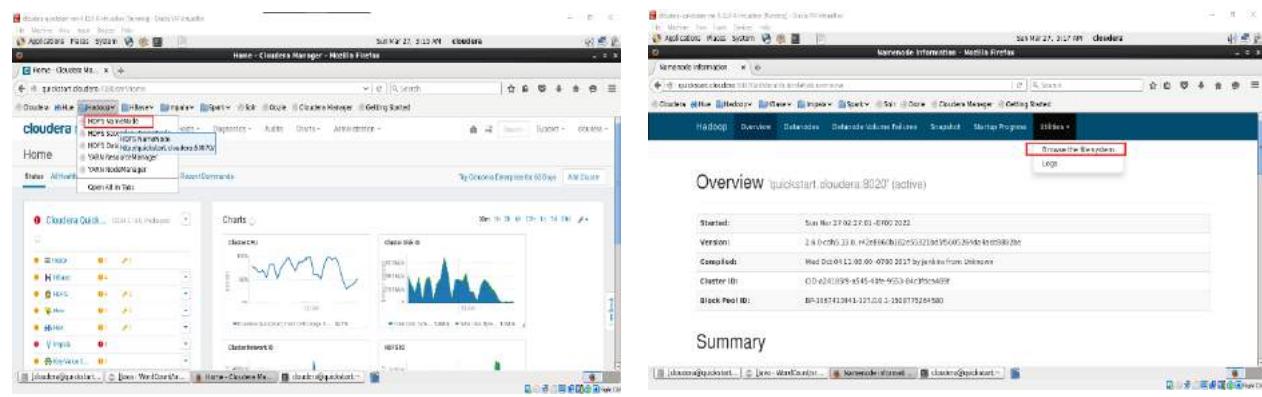


It will give the count of number of times each word has occurred as output.

12. The same file can also be accessed using a browser.

[Browse the Directory by](#)

Hadoop->HDFS Namenode->Utilities ->Browse the file system



Browsing HDFS - Mozilla Firefox

Sun Mar 27, 3:20 AM cloudera

Browsing HDFS - Mozilla Firefox

Sun Mar 27, 3:22 AM cloudera

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

/outputdir Got

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwxrwxrwx	hdfs	supergroup	0 B	Mon Oct 23 09:15:43 -0700 2017	0	0 B	benchmarks
drwxr-xr-x	hbase	supergroup	0 B	Sun Mar 27 02:30:39 -0700 2022	0	0 B	hbase
drwxr-xr-x	hdfs	supergroup	0 B	Sun Mar 27 02:55:49 -0700 2022	0	0 B	inputdir
drwxr-xr-x	hdfs	supergroup	0 B	Thu Mar 17 07:58:45 -0700 2022	0	0 B	newdir
drwxr-xr-x	hdfs	supergroup	0 B	Sun Mar 27 03:08:55 -0700 2022	0	0 B	outputdir
drwxr-xr-x	hdfs	supergroup	0 B	Fri Mar 18 05:36:11 -0700 2022	0	0 B	rjlocal
drwxr-xr-x	solr	solr	0 B	Mon Oct 23 09:18:01 -0700 2017	0	0 B	solr
drwxrwxrwt	hdfs	supergroup	0 B	Fri Mar 25 06:12:06 -0700 2022	0	0 B	tmp
drwxr-xr-x	hdfs	supergroup	0 B	Thu Mar 17 07:20:17 -0700 2022	0	0 B	user
drwxr-xr-x	hdfs	supergroup	0 B	Mon Oct 23 09:17:24 -0700 2017	0	0 B	var

[cloudera@quickstart... Java - WordCount/sr... Browsing HDFS - Mozilla Firefox cloudera@quickstart:~] Sun Mar 27, 3:20 AM cloudera

[cloudera@quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox File Machine View Input Devices Help Applications Places System Browsing HDFS - Mozilla Firefox cloudera@quickstart:~] Sun Mar 27, 3:22 AM cloudera

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

Hadoop Overview Datanodes Snapshot Startup Progress Utilities

Browse Directory

Browsing HDFS - Mozilla Firefox

Sun Mar 27, 3:22 AM cloudera

Browsing HDFS - Mozilla Firefox

Sun Mar 27, 3:22 AM cloudera

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

/outputdir Got

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hdfs	supergroup	0 B	Sun Mar 27 03:08:55 -0700 2022	1	128 MB	SUCCESS
-rw-r--r--	hdfs	supergroup	552 B	Sun Mar 27 03:08:52 -0700 2022	1	128 MB	part-r-00000

Hadoop, 2017.

Now downloading the part-r-00000 file.

Browsing HDFS - Mozilla Firefox

Sun Mar 27, 3:23 AM cloudera

Browsing HDFS - Mozilla Firefox

Sun Mar 27, 3:23 AM cloudera

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

Hadoop Overview Datanodes Snapshot Startup Progress Utilities

File information - part-r-00000

Download

Opening part-r-00000

You have chosen to open:
part-r-00000
which is: BIN file (552 bytes)
from: http://quickstart.cloudera:50070

Would you like to save this file?

Cancel Save File

Availability:
quickstart.cloudera

Block Size Name

128 MB SUCCESS

128 MB part-r-00000

Browsing HDFS - Mozilla Firefox

Sun Mar 27, 3:25 AM cloudera

part-r-00000 (~:/Downloads) - gedit

File Edit View Search Tools Documents Help

Open Save Undo Redo

part-r-00000

```
laboris 1
latis 1
laborum 1
magna 1
minim 1
nullit 1
nisi 1
nostrud 1
nulla 1
occaecat 1
officia 1
perferitur 1
president 1
qui 1
quis 1
reprehenderit 1
tempor 1
sint 1
sit 1
sunt 1
tempor 1
utemco 1
ut 2
velit 1
veniam 1
voluptate 1
```

Plain Text Tab Width: 8 Ln 1, Col 1

For every execution of this program we need to delete the output directory or give a new name to the output directory every time.

Inside the part-r-00000 file it will have the same output as we are getting after executing using command
hadoop jar /home/cloudera/WordCount.jar WordCount /inputdir/abc /op1

Sun Mar 27, 3:32 AM cloudera

```
cloudera@quickstart:~$ hdfs dfs -ls /
Found 11 items
drwxrwxrwx - hdfs supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup          0 2022-03-27 02:30 /hbase
drwxr-xr-x - hdfs supergroup          0 2022-03-27 02:55 /inputdir
drwxr-xr-x - hdfs supergroup          0 2022-03-17 07:58 /newdir
drwxr-xr-x - hdfs supergroup          0 2022-03-27 03:31 /op1
drwxr-xr-x - hdfs supergroup          0 2022-03-27 03:08 /outputdir
drwxr-xr-x - solr   supergroup          0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup          0 2022-03-25 08:12 /tmp
drwxr-xr-x - hdfs supergroup          0 2022-03-17 07:20 /user
drwxr-xr-x - hdfs supergroup          0 2017-10-23 09:17 /var
[cloudera@quickstart ~]$ hdfs dfs -ls /op1
Found 2 items
-rw-r--r-- 1 hdfs supergroup          0 2022-03-27 03:31 /op1/_SUCCESS
-rw-r--r-- 1 hdfs supergroup 552 2022-03-27 03:31 /op1/part-r-00000
[cloudera@quickstart ~]$
```

Sun Mar 27, 3:32 AM cloudera

```
cloudera@quickstart:~$ hadoop jar /home/cloudera/WordCount.jar wordCount /inputdir/abc /op1
22/03/27 03:28:54 INFO client.RMProxy: Connecting to ResourceManager at quickstart.cloudera:19882
22/03/27 03:28:54 WARN mapreduce.JobSubmitter: No implementation found for interface: org.apache.hadoop.mapreduce.Job
22/03/27 03:28:54 WARN mapreduce.JobSubmitter: Implement the Tool Interface and execute your application with ToolRunner to remedy this.
22/03/27 03:29:00 INFO input.FileInputFormat: Total input paths to process : 1
22/03/27 03:29:00 INFO mapreduce.Job: Number of reduce tasks determined: 1
22/03/27 03:29:04 INFO mapreduce.Job: Submitting token for job: job_16048373372244_0002
22/03/27 03:29:10 INFO impl.YarnClientImpl: Submitted application application_16048373372244_0002
22/03/27 03:29:10 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/
22/03/27 03:29:10 INFO mapreduce.Job: Running job: job_16048373372244_0002
22/03/27 03:30:12 INFO mapreduce.Job: Job job_16048373372244_0002 running in uber mode : false
22/03/27 03:30:12 INFO mapreduce.Job: 0% complete
22/03/27 03:30:12 INFO mapreduce.Job: 0% reduce
22/03/27 03:31:11 INFO mapreduce.Job: map 100% reduce 0%
22/03/27 03:31:11 INFO mapreduce.Job: map 100% reduce 100%
22/03/27 03:31:13 INFO mapreduce.Job: Job job_16048373372244_0002 completed successfully
22/03/27 03:31:13 INFO mapreduce.Job: Counters: 49
File System Counters
  File system bytes read=0
  File system bytes read=0
  File: Number of bytes written=295227
  File: Number of read operations=8
  File: Number of large read operations=0

[cloudera@quick... Java - WordCo... Browsing HDFS... cloudera@quick... part-r-00000 (...)

Sun Mar 27, 3:32 AM cloudera
```

Sun Mar 27, 3:32 AM cloudera

```
cloudera@quickstart:~$ hadoop jar /home/cloudera/WordCount.jar wordCount /inputdir/abc /op1
22/03/27 03:28:54 INFO client.RMProxy: Connecting to ResourceManager at quickstart.cloudera:19882
22/03/27 03:28:54 WARN mapreduce.JobSubmitter: No implementation found for interface: org.apache.hadoop.mapreduce.Job
22/03/27 03:28:54 WARN mapreduce.JobSubmitter: Implement the Tool Interface and execute your application with ToolRunner to remedy this.
22/03/27 03:29:00 INFO input.FileInputFormat: Total input paths to process : 1
22/03/27 03:29:00 INFO mapreduce.Job: Number of reduce tasks determined: 1
22/03/27 03:29:04 INFO mapreduce.Job: Submitting token for job: job_16048373372244_0002
22/03/27 03:29:10 INFO impl.YarnClientImpl: Submitted application application_16048373372244_0002
22/03/27 03:29:10 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/
22/03/27 03:29:10 INFO mapreduce.Job: Running job: job_16048373372244_0002
22/03/27 03:30:12 INFO mapreduce.Job: Job job_16048373372244_0002 running in uber mode : false
22/03/27 03:30:12 INFO mapreduce.Job: 0% complete
22/03/27 03:30:12 INFO mapreduce.Job: 0% reduce
22/03/27 03:31:11 INFO mapreduce.Job: map 100% reduce 0%
22/03/27 03:31:11 INFO mapreduce.Job: map 100% reduce 100%
22/03/27 03:31:13 INFO mapreduce.Job: Job job_16048373372244_0002 completed successfully
22/03/27 03:31:13 INFO mapreduce.Job: Counters: 49
Map-Reduce Framework
  Map input records=1
  Map output records=0
  Map output bytes=222
  Map output materialized bytes=656
  Input split Bytes=109
  Combine input records=0
  Combine output records=0
  Reduce input groups=64
  Reduce shuffle bytes=450
  Redundant bytes=0
  Reduced output records=64
  Spilled Records=130
  Shuffled Maps=64
  Failed Shuffles=0
  Merged Map output=1
  GC time spent (ms)=267
  CPU time spent (ms)=479
  Physical memory (bytes) snapshot=244666368
  Virtual memory (bytes) snapshot=1481925998
  Total committed heap usage (bytes)=969953280

[cloudera@quick... Java - WordCo... Browsing HDFS... cloudera@quick... part-r-00000 (...)

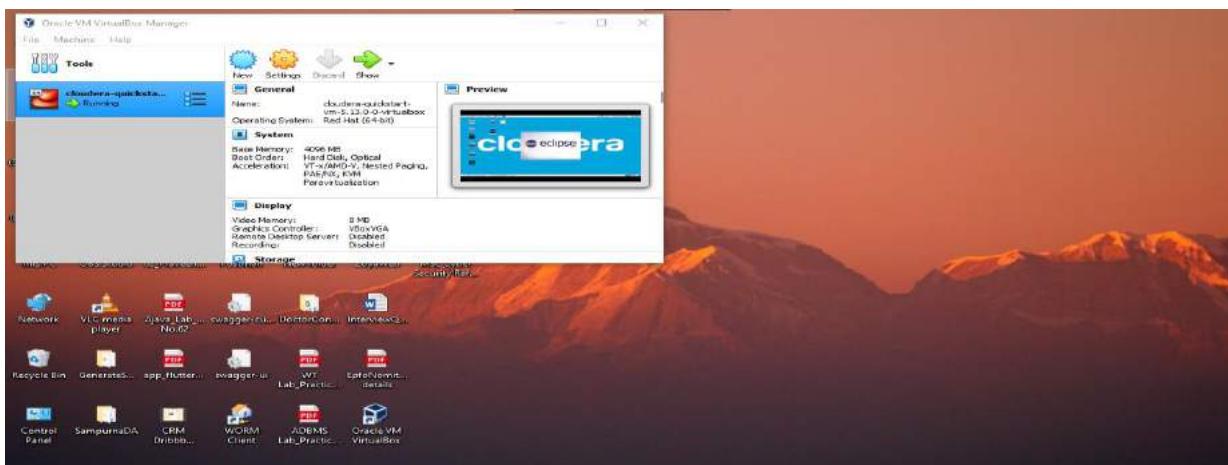
Sun Mar 27, 3:32 AM cloudera
```

Aim: To determine maximum temperature using Hadoop MapReduce**Dataset – temperature.txt**

```
0067011990999991950051507004+68750+023550FM-12+038299999V0203301N00671220001CN9999999N9+00001+999999999999
0043011990999991950051512004+68750+023550FM-12+038299999V0203201N00671220001CN9999999N9+00221+999999999999
0043011990999991950051518004+68750+023550FM-12+038299999V0203201N00261220001CN9999999N9-00111+999999999999
0043012650999991949032412004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9+01111+999999999999
0043012650999991949032418004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9+00781+999999999999
```

Steps to determine maximum temperature using Hadoop MapReduce in Cloudera:

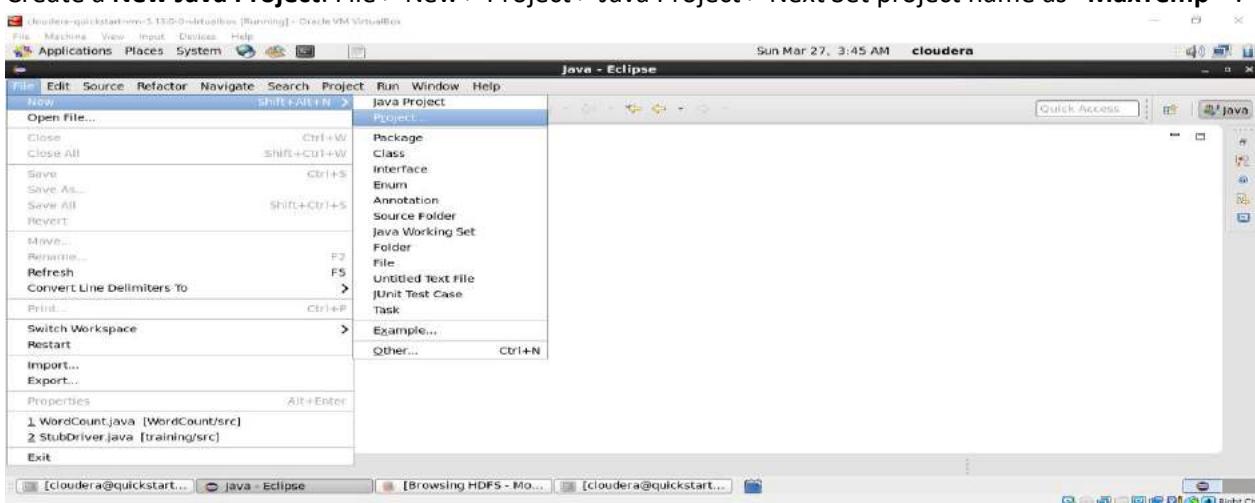
1. Open VirtualBox and then start Cloudera VM

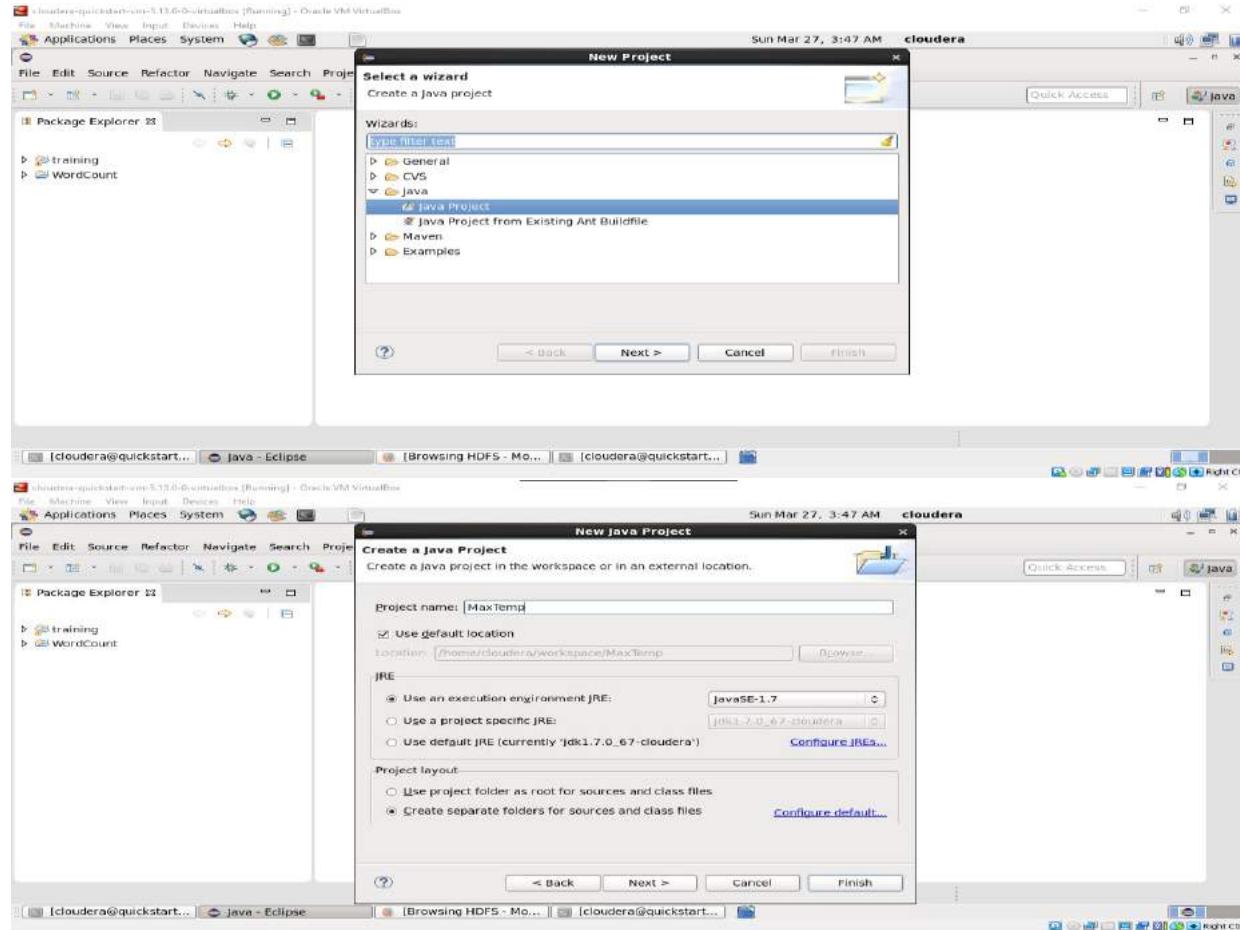


2. Open Eclipse



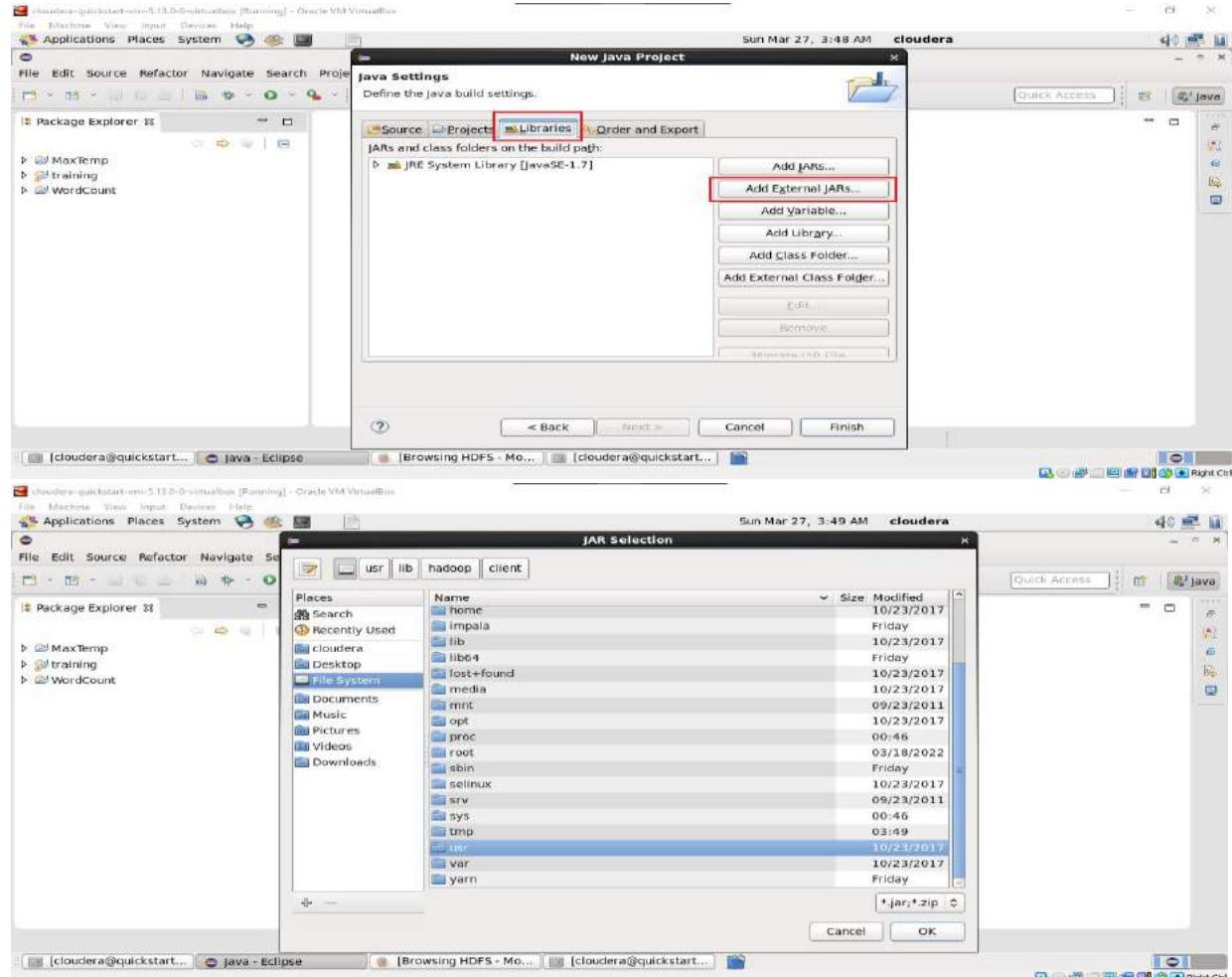
3. Create a New Java Project: File > New > Project > Java Project > Next Set project name as "MaxTemp".

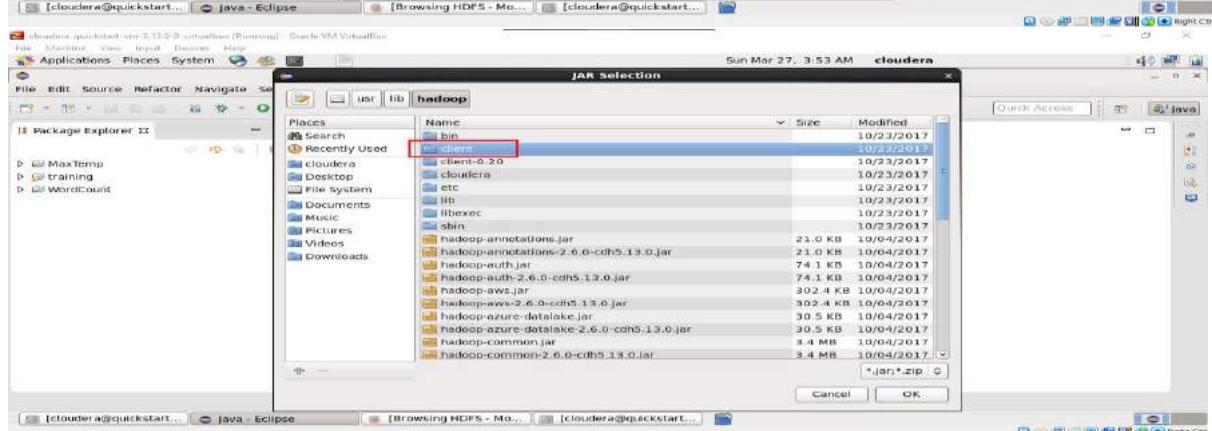
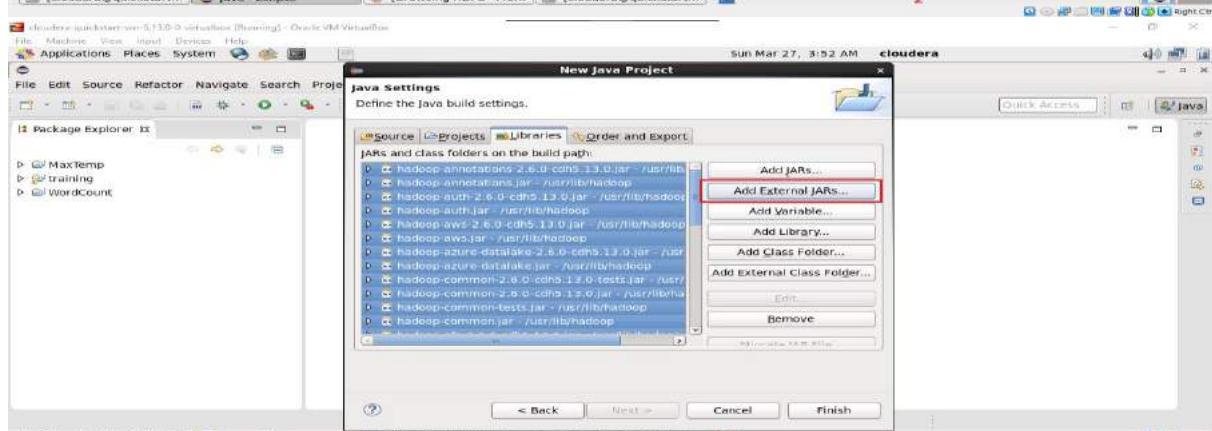
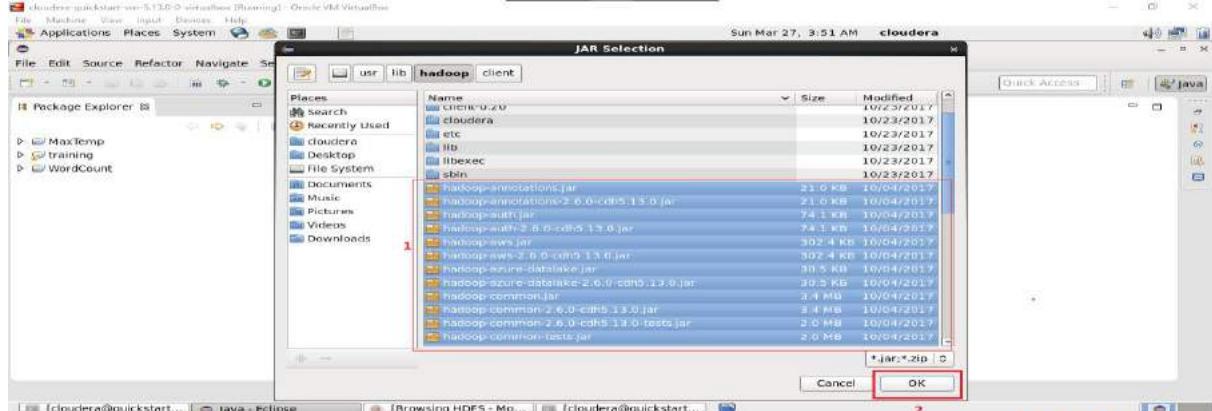
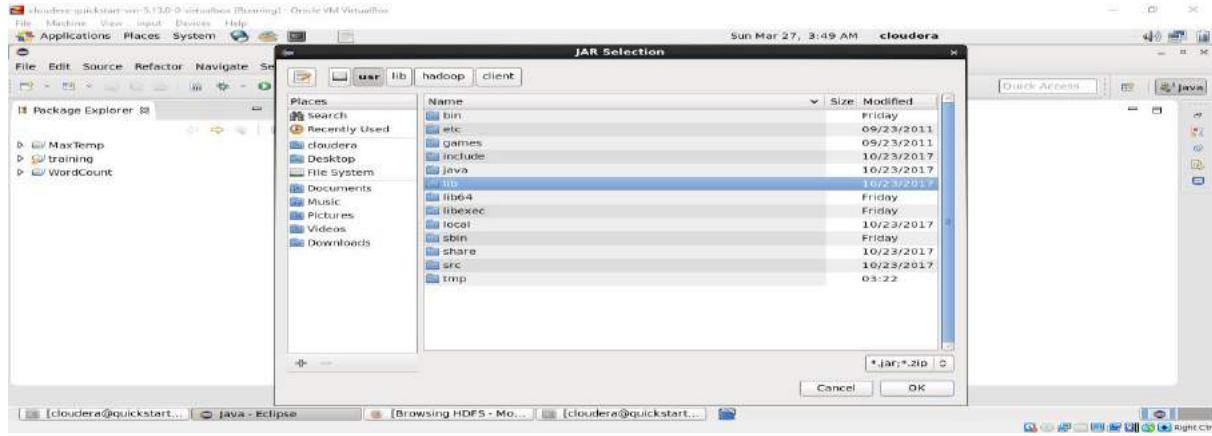




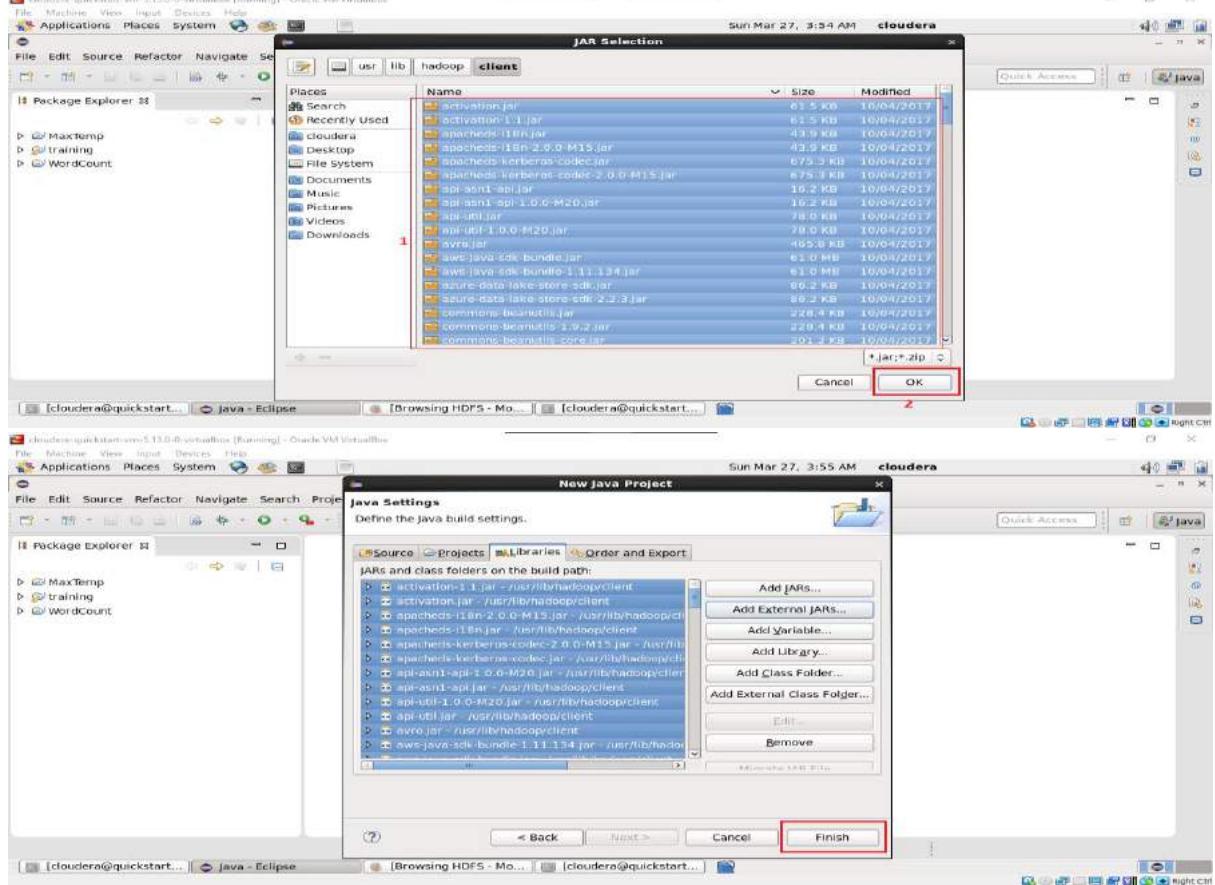
4. Adding the hadoop libraries to the project click on:

Libraries > Add External JARs > File System > usr > lib > hadoop > client > select all JAR files > OK > Finish.





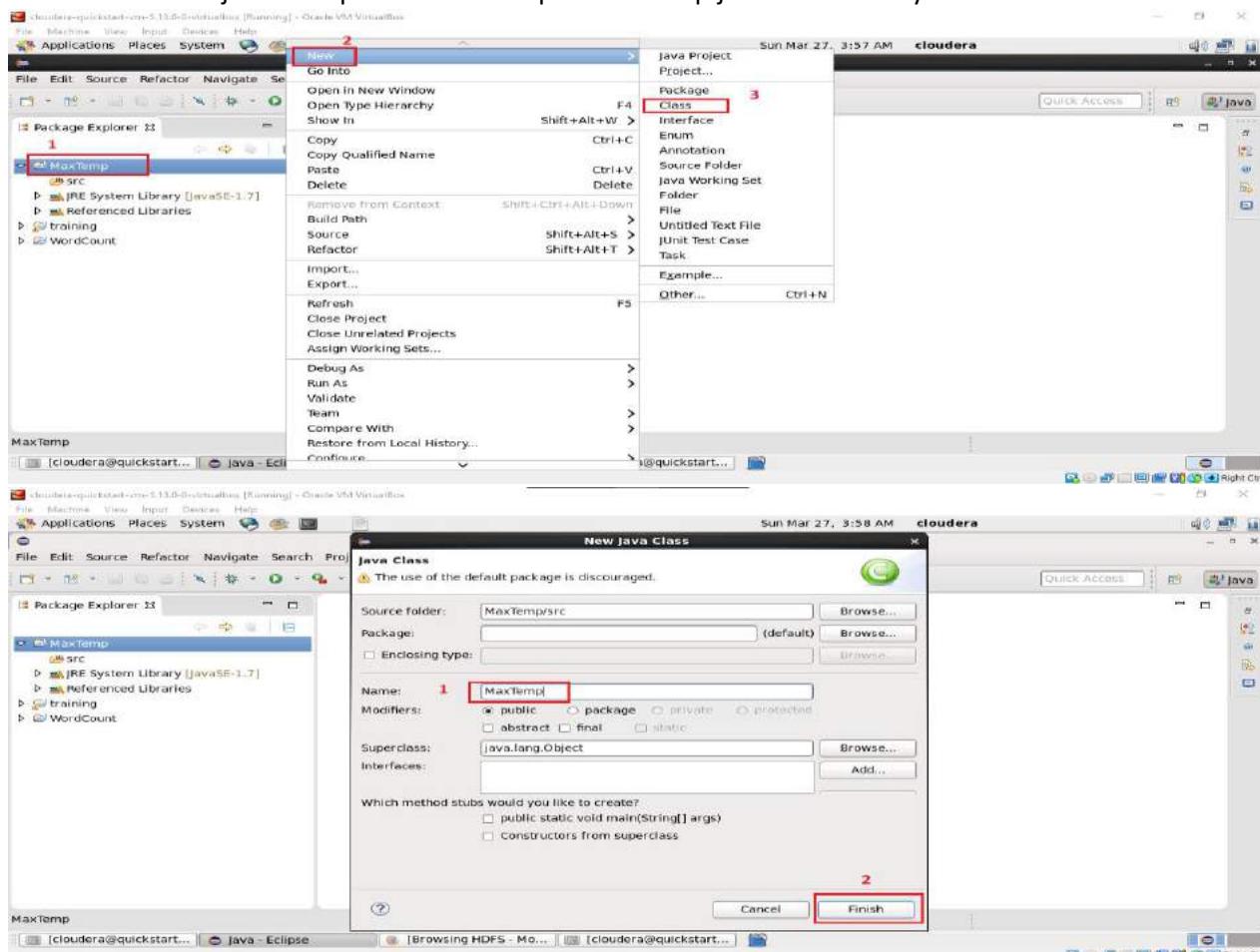
Name: Pavan Yadav



5. Right click on the name of project

"MaxTemp" > New > Class (Don't write anything for package) > class name "MaxTemp" > Finish

Then WordCount.java will open if not then open MaxTemp.java file manually



6. Source Code:

```
1 //Packages
2
3+import java.io.IOException;
18
19 public class MaxTemp {
20     //Map Logic
21+    public static class MaxTemperatureMapper extends Mapper < LongWritable, Text, Text, IntWritable > {
39
40     //Reducer
41+    public static class MaxTemperatureReducer extends Reducer < Text, IntWritable, Text, IntWritable > {
50         public void reduce(Text key, Iterable < IntWritable > values, Context context) throws IOException, InterruptedException {
51             int maxvalue = Integer.MIN_VALUE;
52+            public static void main(String[] args) throws Exception {
70 }
71 }
```

```
//Packages

import java.io.IOException;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.Text;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;

import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.conf.Configuration;

public class MaxTemp {
    //Map Logic
    public static class MaxTemperatureMapper extends Mapper < LongWritable, Text, Text,
    IntWritable > {

        public static final int MISSING = 9999;

        public void map(LongWritable key, Text value, Context context) throws IOException,
        InterruptedException {
            String line = value.toString();
            String year = line.substring(15, 19);
            int airtemp;
            if (line.charAt(87) == '+') {
                airtemp = Integer.parseInt(line.substring(88, 92));
            } else
                airtemp = Integer.parseInt(line.substring(87, 92));
            String q = line.substring(92, 93);
            if (airtemp != MISSING && q.matches("[01459]")) {
                context.write(new Text(year), new IntWritable(airtemp));
            }
        }
    }

    //Reducer
    public static class MaxTemperatureReducer extends Reducer < Text, IntWritable, Text,
    IntWritable > {
```

```

        public void reduce(Text key, Iterable < IntWritable > values, Context context) throws
IOException, InterruptedException {
    int maxvalue = Integer.MIN_VALUE;
    for (IntWritable value: values) {
        maxvalue = Math.max(maxvalue, value.get());
    }
    context.write(key, new IntWritable(maxvalue));
}

//Main Function
public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();

    Job job = Job.getInstance(conf, "weather example");
    job.setJarByClass(MaxTemp.class);
    job.setMapperClass(MaxTemperatureMapper.class);

    job.setReducerClass(MaxTemperatureReducer.class);

    job.setInputFormatClassTextInputFormat.class);

    job.setOutputFormatClass(TextOutputFormat.class);
    job.setOutputKeyClass(Text.class);
    job.setOutputValueClass(IntWritable.class);
    FileInputFormat.addInputPath(job, new Path(args[0]));
    FileOutputFormat.setOutputPath(job, new Path(args[1]));
    System.exit(job.waitForCompletion(true) ? 0 : 1);
}
}
}

```

```

1 //Packages
2
3 import java.io.IOException;
4 import org.apache.hadoop.fs.Path;
5 import org.apache.hadoop.io.LongWritable;
6
7 import org.apache.hadoop.io.Text;
8 import org.apache.hadoop.io.IntWritable;
9 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
10 import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
11 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
12 import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
13
14 import org.apache.hadoop.mapreduce.Job;
15 import org.apache.hadoop.mapreduce.Mapper;
16 import org.apache.hadoop.mapreduce.Reducer;
17 import org.apache.hadoop.conf.Configuration;
18
19 /**
20 * Max Temperature Mapper
21 */
22 public static class MaxTemperatureMapper extends Mapper < LongWritable, Text, Text, IntWritable > {
23
24     public static final int MISSING = 9999;
25
26     public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {
27         String line = value.toString();
28         String year = line.substring(15, 19);
29         int airtemp;
30         if (line.charAt(87) == '+') {
31             airtemp = Integer.parseInt(line.substring(88, 92));
32         } else {
33             airtemp = Integer.parseInt(line.substring(87, 92));
34         }
35         String q = line.substring(92, 93);
36         if (airtemp != MISSING && q.matches("[01459]")) {
37             context.write(new Text(year), new IntWritable(airtemp));
38         }
39     }
40 }

```

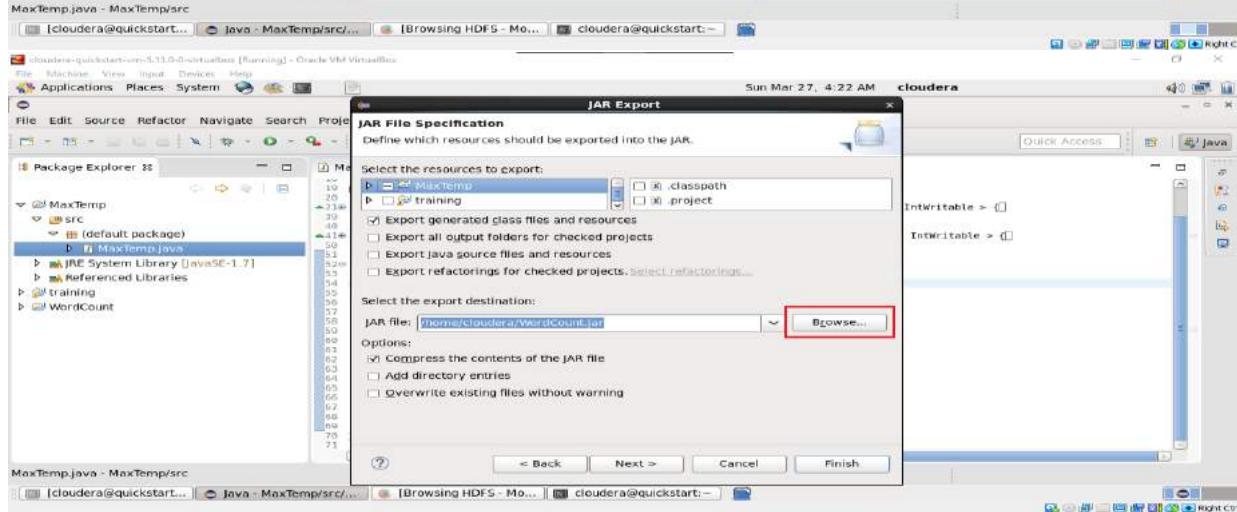
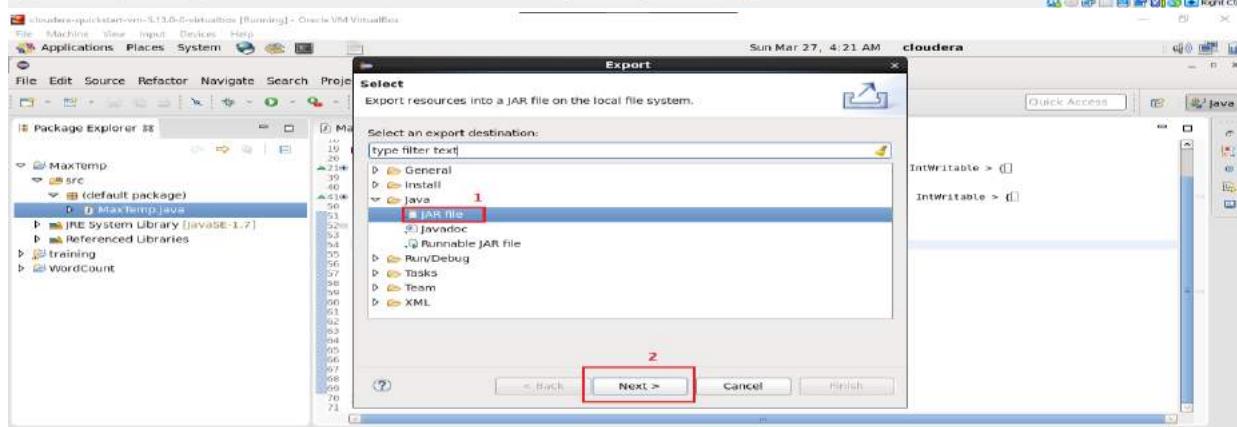
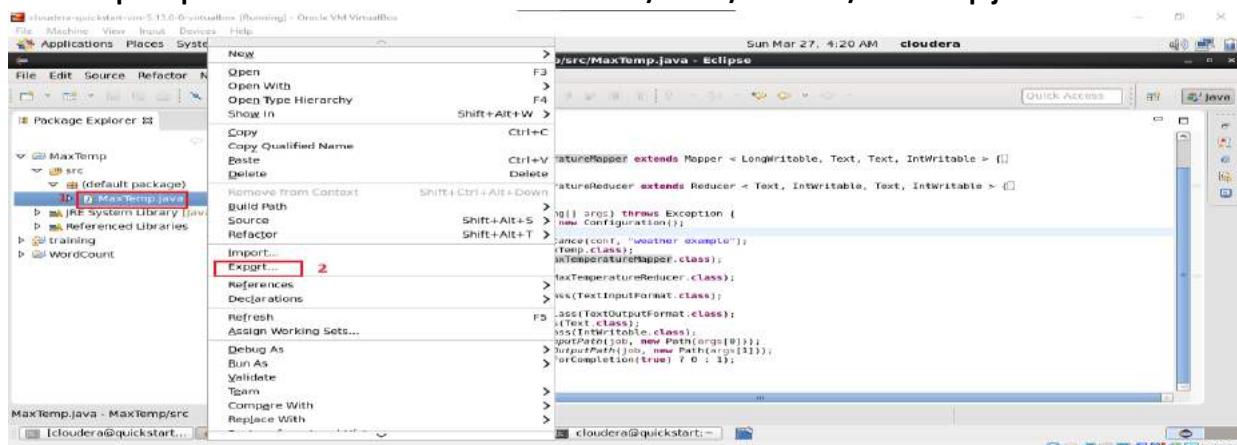
```

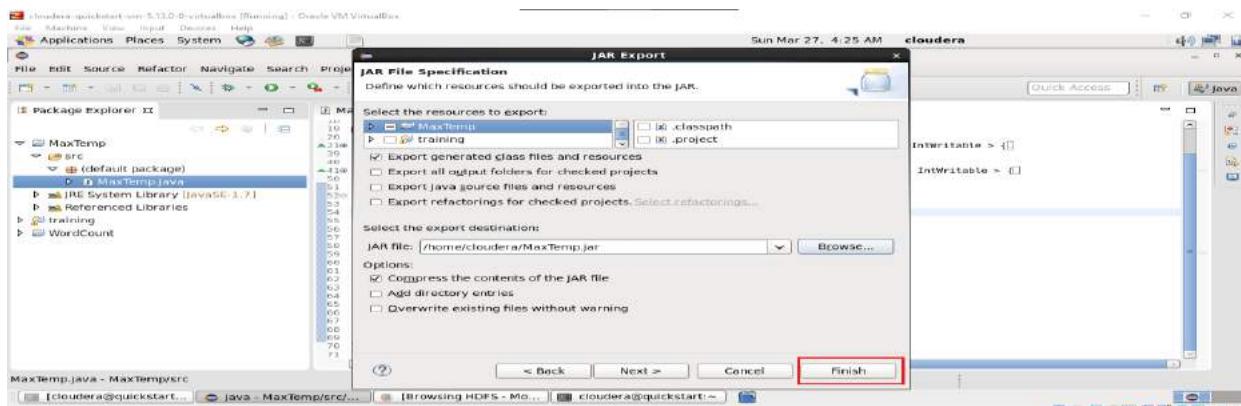
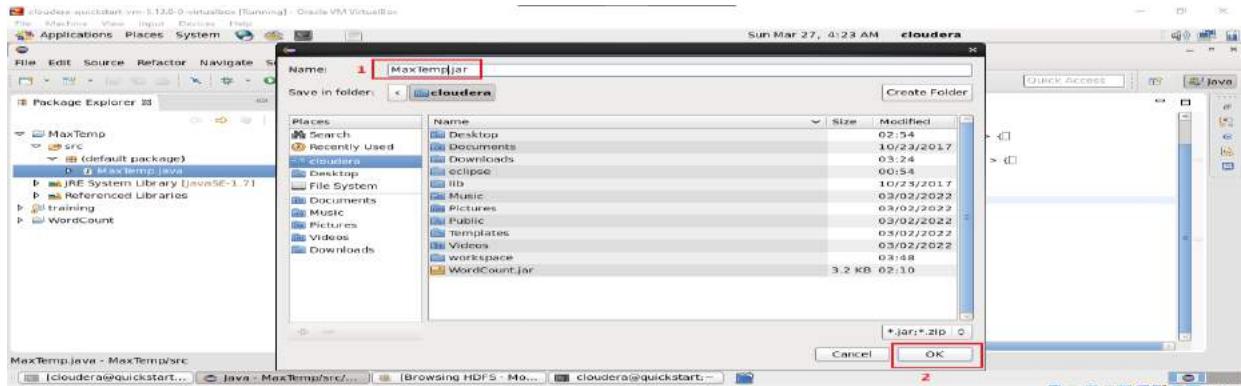
40 //Reducer
41 public static class MaxTemperatureReducer extends Reducer < Text, IntWritable, Text, IntWritable > {
42     public void reduce(Text key, Iterable < IntWritable > values, Context context) throws IOException, InterruptedException {
43         int maxvalue = Integer.MIN_VALUE;
44         for (IntWritable value: values) {
45             maxvalue = Math.max(maxvalue, value.get());
46         }
47         context.write(key, new IntWritable(maxvalue));
48     }
49 }
50
51 //Main Function
52 public static void main(String[] args) throws Exception {
53     Configuration conf = new Configuration();
54
55     Job job = Job.getInstance(conf, "weather example");
56     job.setJarByClass(MaxTemp.class);
57     job.setMapperClass(MaxTemperatureMapper.class);
58
59     job.setReducerClass(MaxTemperatureReducer.class);
60
61     job.setInputFormatClass(TextInputFormat.class);
62
63     job.setOutputFormatClass(TextOutputFormat.class);
64     job.setOutputKeyClass(Text.class);
65     job.setOutputValueClass(IntWritable.class);
66     TextInputFormat.addInputPath(job, new Path(args[0]));
67     FileOutputFormat.setOutputPath(job, new Path(args[1]));
68     System.exit(job.waitForCompletion(true) ? 0 : 1);
69 }

```

7. Right click on the project name

MaxTemp> Export > Java > JAR File > Next > Jar file “/home/cloudera/MaxTemp.jar”> Finish.





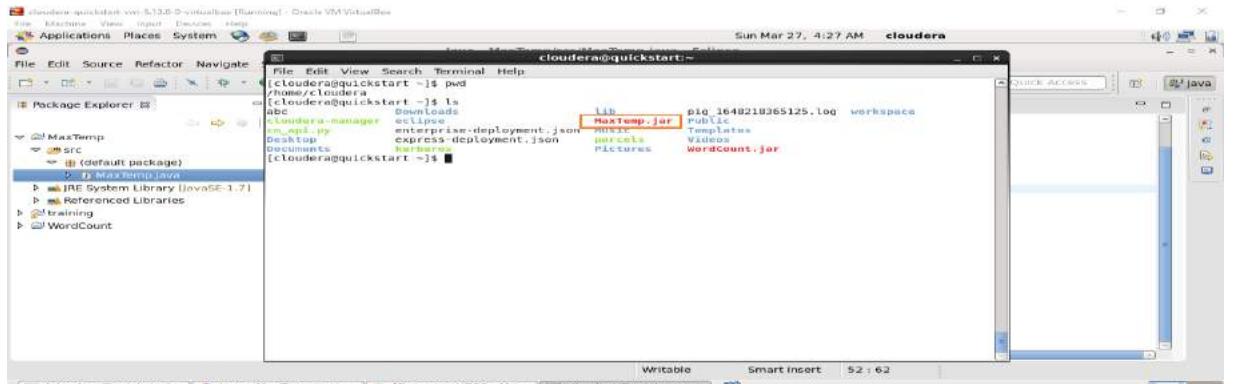
Verify jar file from terminal by using Open terminal & type “ ls ” There it will show MaxTemp.jar

Check current working directory

pwd

Check the list

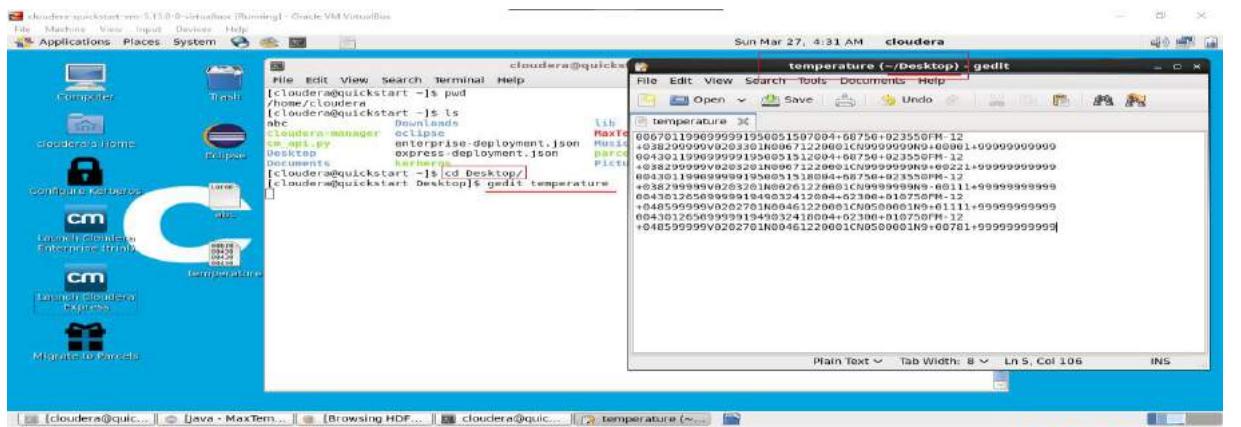
Is



8. We need to create an **input file** in local file system (On Desktop)

Creating an input file named as “**temperature**”.

Paste the **DataSet**



Here listing all the directory present in

hdfs using hdfs dfs -ls /

```
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 11 items
drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup 0 2022-03-27 02:30 /hbase
drwxr-xr-x - hdfs supergroup 0 2022-03-27 02:55 /inputdir
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:58 /newdir
drwxr-xr-x - hdfs supergroup 0 2022-03-27 03:31 /op1
drwxr-xr-x - hdfs supergroup 0 2022-03-27 03:08 /outputdir
drwxr-xr-x - hdfs supergroup 0 2022-03-18 05:36 /rjclocal
drwxr-xr-x - solr solr 0 2017-10-23 09:18 /solr
drwxrwxrwt - hdfs supergroup 0 2022-03-25 08:12 /tmp
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:20 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /var
[cloudera@quickstart Desktop]$
```

9. Now we have to move this input file to hdfs. For this we create a direcory on hdfs using command

hdfs dfs -mkdir /tempip

```
[cloudera@quickstart Desktop]$ hdfs dfs -mkdir /tempip
[cloudera@quickstart Desktop]$
```

If getting any permission related error than execute below command

export HADOOP_USER_NAME=hdfs

Then we can verify whether this directory is created or not using ls command

hdfs dfs -ls /

```
[cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 12 items
drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup 0 2022-03-27 02:30 /hbase
drwxr-xr-x - hdfs supergroup 0 2022-03-27 02:55 /inputdir
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:58 /newdir
drwxr-xr-x - hdfs supergroup 0 2022-03-27 03:31 /op1
drwxr-xr-x - hdfs supergroup 0 2022-03-27 03:08 /outputdir
drwxr-xr-x - hdfs supergroup 0 2022-03-18 05:36 /rjclocal
drwxr-xr-x - solr solr 0 2017-10-23 09:18 /solr
drwxr-xr-x - hdfs supergroup 0 2022-03-27 04:37 /tempip
drwxrwxrwt - hdfs supergroup 0 2022-03-25 08:12 /tmp
drwxr-xr-x - hdfs supergroup 0 2022-03-17 07:20 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /var
[cloudera@quickstart Desktop]$
```

Move the input file to this directory created in hdfs by using either **put** command or **copyFromLocal** command

hdfs dfs -put /home/cloudera/Desktop/temperature /tempip/

```
[cloudera@quickstart Desktop]$ hdfs dfs -put /home/cloudera/Desktop/temperature /tempip/
[cloudera@quickstart Desktop]$
```

Now checking whether the “temperature” present in /tempip directory of hdfs or not using command

hdfs dfs -ls /tempip

```
[cloudera@quickstart Desktop]$ hdfs dfs -ls /tempip
Found 1 items
-rw-r--r-- 1 hdfs supergroup 530 2022-03-27 04:37 /tempip/temperature
[cloudera@quickstart Desktop]$
```

As we can see “temperature” file is present in / tempip directory of hdfs. Now we will see the content of this file using command

hdfs dfs -cat /tempip/temperature

```
[cloudera@quickstart Desktop]$ hdfs dfs -cat /tempip/temperature
0067011990999991950051507004+68750+023550FM-12+038299999V0203301N00671220001CN9999999N9+00001+99999999
999
0043011990999991950051512004+68750+023550FM-12+038299999V0203201N00671220001CN9999999N9+00221+99999999
999
0043011990999991950051518004+68750+023550FM-12+038299999V0203201N00261220001CN9999999N9-00111+99999999
999
0043012650999991949032412004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9+01111+99999999
999
0043012650999991949032418004+62300+010750FM-12+048599999V0202701N00461220001CN0500001N9+00781+99999999
999
[cloudera@quickstart Desktop]$
```

10. Running Mapreduce Program on Hadoop,

Syntax : **hadoop jar jarFileName.jar ClassName /InputFileAddress /outputdir**

hadoop jar /home/cloudera/MaxTemp.jar MaxTemp /tempip/temperature /tempop1

```
[cloudera@quickstart Desktop]$ hadoop jar /home/cloudera/MaxTemp.jar MaxTemp /tempip/temperature /tempop
22/03/27 05:53:49 INFO client.RMProxy: Connecting to ResourceManager at quickstart.cloudera/10.0.2.15:8032
22/03/27 05:53:56 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
```

Map-Reduce Framework

```
cloudera@quickstart:~/Desktop$ hadoop jar /home/cloudera/MaxTemp.jar MaxTemp /tempip/temperature /tempop
22/03/27 05:54:14 INFO client.RMProxy: Connecting to ResourceManager at quickstart.cloudera/10.0.2.15:8032
22/03/27 05:54:17 INFO mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
22/03/27 05:54:17 INFO mapreduce.JobResourceUploader: Datanode connection attempt 1 failed; trying again: 10.0.2.15:54310
22/03/27 05:54:17 INFO mapreduce.JobResourceUploader: Datanode connection attempt 2 succeeded
22/03/27 05:54:17 INFO mapreduce.JobResourceUploader: Submittting token for job job_1648373372244_0003
22/03/27 05:54:17 INFO mapreduce.JobResourceUploader: Submittting application application_1648373372244_0003
22/03/27 05:54:17 INFO mapreduce.JobResourceUploader: Job job_1648373372244_0003 successfully
22/03/27 05:54:17 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1648373372244_0003
22/03/27 05:54:17 INFO mapreduce.Job: Counters
FILE: Number of bytes read=62
FILE: Number of bytes written=294943
FILE: Number of small read operations=1
FILE: Number of large read operations=8
FILE: Number of small write operations=6
HDFS: Number of bytes read=6484
HDFS: Number of bytes written=17
HDFS: Number of small read operations=1
HDFS: Number of large read operations=1
HDFS: Number of write operations=2
Job Counters
  FILE: Number of bytes read=62
  FILE: Number of bytes written=294943
  FILE: Number of small read operations=1
  FILE: Number of large read operations=8
  FILE: Number of small write operations=6
  HDFS: Number of bytes read=6484
  HDFS: Number of bytes written=17
  HDFS: Number of small read operations=1
  HDFS: Number of large read operations=1
  HDFS: Number of write operations=2
Job Counters
```

```
cloudera@quickstart:~/Desktop$ hadoop jar /home/cloudera/MaxTemp.jar MaxTemp /tempip/temperature /tempop
22/03/27 05:55:07 INFO mapreduce.Job: Job job_1648373372244_0003 running in uber mode : false
22/03/27 05:55:07 INFO mapreduce.Job: map 100% reduce 0%
22/03/27 05:55:07 INFO mapreduce.Job: map 100% reduce 100%
22/03/27 05:55:28 INFO mapreduce.Job: map 100% reduce 100%
22/03/27 05:55:28 INFO mapreduce.Job: Job job_1648373372244_0003 completed successfully
22/03/27 05:56:24 INFO mapreduce.Job: Counters
  FILE: Number of bytes read=62
  FILE: Number of bytes written=294943
  FILE: Number of small read operations=1
  FILE: Number of large read operations=8
  FILE: Number of small write operations=6
  HDFS: Number of bytes read=6484
  HDFS: Number of bytes written=17
  HDFS: Number of small read operations=1
  HDFS: Number of large read operations=1
  HDFS: Number of write operations=2
Job Counters
  Map output bytes=5
  Map output materialized bytes=50
  Map input records=5
  Combine input records=0
  Combiner input records=0
  Reduce input groups=2
  Reduce shuffle bytes=50
  Reduce input bytes=50
  Reduce output records=2
  Spilled Records=10
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ms)=355
  CPU time spent (ms)=100
  Physical memory (bytes) snapshot=246063104
  Virtual memory (bytes) snapshot=147995216
  Total committed heap usage (bytes)=95944704
Shuffle Errors
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONGReducer=0
  File Input Format Counters
    bytes read=6484
    File Output Format Counters
      bytes written=17
[cloudera@quickstart desktop]$
```

11. Then we can verify the content of outputdir directory and in that part-r file has the actual output by using the command `Hdfs dfs -cat /outputdir/part-r-00000` This will give us final output. The same file can also be accessed using a browser. For every execution of this program we need to delete the output directory or give a new name to the output directory every time. 1st we are checking whether the outputdir directory is created in hdfs or not using command

hdfs dfs -ls /

```
cloudera@quickstart Desktop]$ hdfs dfs -ls /
Found 13 items
drwxrwxrwx  - hdfs  supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x  - hbase supergroup          0 2022-03-27 05:48 /hbase
drwxr-xr-x  - hdfs  supergroup          0 2022-03-27 02:55 /inputdir
drwxr-xr-x  - hdfs  supergroup          0 2022-03-17 07:58 /newdir
drwxr-xr-x  - hdfs  supergroup          0 2022-03-27 03:31 /op1
drwxr-xr-x  - hdfs  supergroup          0 2022-03-27 03:08 /outputdir
drwxr-xr-x  - hdfs  supergroup          0 2022-03-18 05:36 /rjclocal
drwxr-xr-x  - solr   supergroup          0 2017-10-23 09:18 /solr
drwxr-xr-x  - hdfs  supergroup          0 2022-03-27 04:37 /tempip
drwxr-xr-x  - hdfs  supergroup          0 2022-03-27 05:56 /tempop
drwxrwxrwx  - hdfs  supergroup          0 2022-03-25 08:12 /tmp
drwxr-xr-x  - hdfs  supergroup          0 2022-03-17 07:20 /user
drwxr-xr-x  - hdfs  supergroup          0 2017-10-23 09:17 /var
[cloudera@quickstart Desktop]$
```

Now let's check what we have inside this **tempop** directory using command as

hdfs dfs -ls /tempop

```
cloudera@quickstart Desktop]$ hdfs dfs -ls /tempop
Found 2 items
-rw-r--r--  1 hdfs supergroup          0 2022-03-27 05:56 /tempop/_SUCCESS
-rw-r--r--  1 hdfs supergroup        17 2022-03-27 05:56 /tempop/part-r-00000
[cloudera@quickstart Desktop]$
```

Now we want to read the content of the part-r-00000 file which present inside the **tempop** using command

hdfs dfs -cat /tempop/part-r-00000

```
cloudera@quickstart Desktop]$ hdfs dfs -cat /tempop/part-r-00000
1949 111
1950 22
[cloudera@quickstart Desktop]$
```

So the maximum temperature for the year 1949 is 111 and for the year 1950 is 22.

12. The same file can also be accessed using a browser.

Browse the Directory by

Hadoop->HDFS Namenode->Utilities ->Browse the file system

The screenshot shows the Cloudera Manager interface with the 'HDFS' tab selected. The left sidebar has 'HDFS DataNodes' highlighted. The main area displays various HDFS metrics and status information.

This screenshot shows the 'Namenode Information' page for the 'quickstart.cloudera:8020' namenode. It provides details like the start time, version, and cluster ID.

The 'Browsing HDFS' page lists files in the '/tmp' directory. A new directory named 'tempop' is shown, with a red box highlighting it and the message 'Created dir on terminal'.

The 'tempop' directory listing shows two files: 'SUCCESS' and 'part-r-00000'. A red box highlights the 'part-r-00000' file.

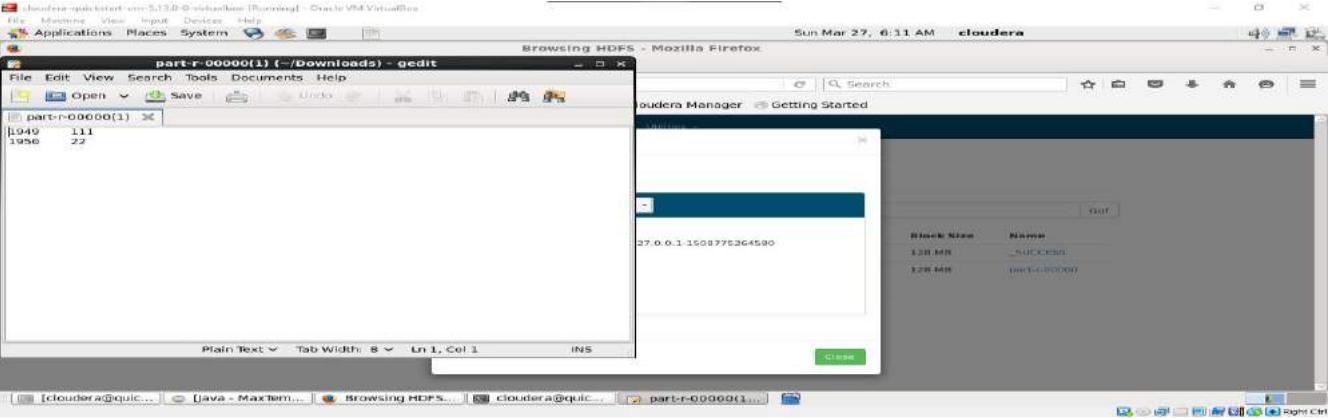
The 'Browse Directory' page shows the same file listing. A red box highlights the 'part-r-00000' file again.

Now downloading the part-r-00000 file.

A modal dialog titled 'File Information - part-r-00000' is displayed. It shows the file path and a 'Download' button. A red box highlights the 'Download' button.

An 'Opening part-r-00000' dialog is shown, asking if the user wants to save the file. A red box highlights the 'Save File' button.

The 'Browse Directory' page now shows the 'part-r-00000' file has been saved successfully.



For every execution of this program we need to delete the output directory or give a new name to the output directory every time.

Inside the part-r-00000 file it will have the same output as we are getting after executing using command.

hadoop jar /home/cloudera/MaximumTemp.jar MaxTemp /tempip/temperature /tempop1

Aim: A- Installing MongoDB

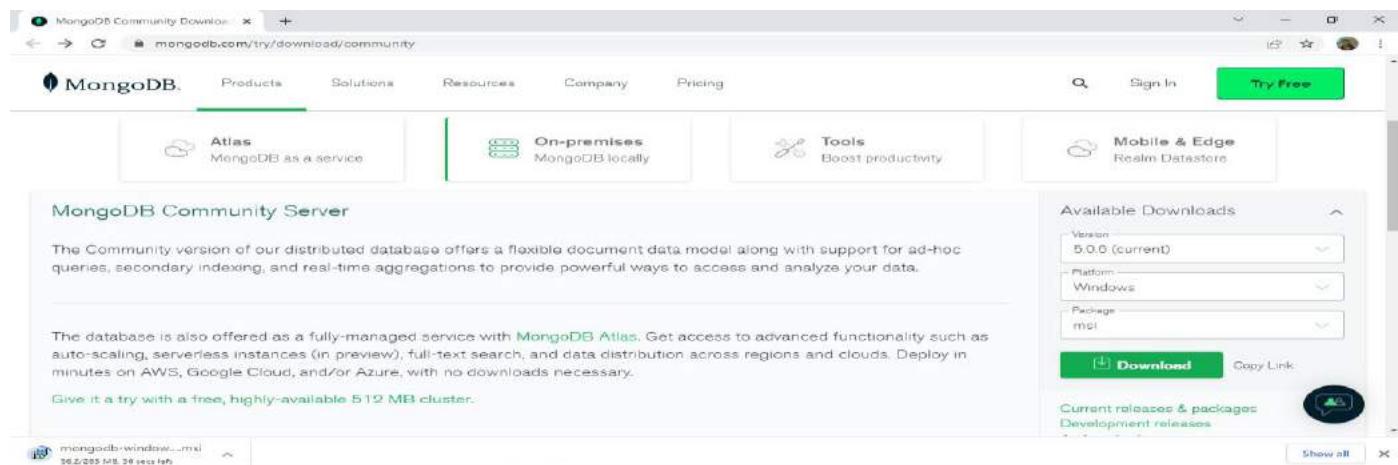
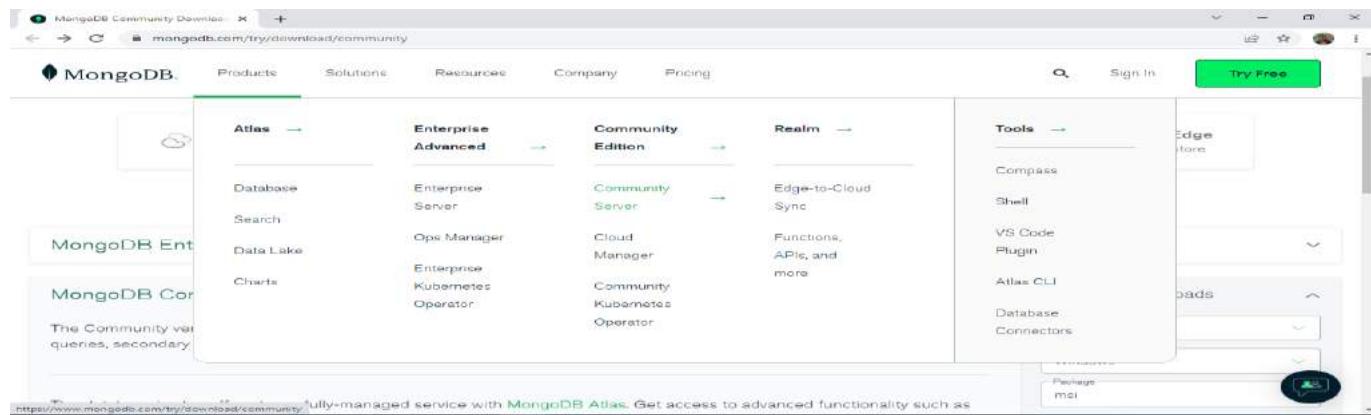
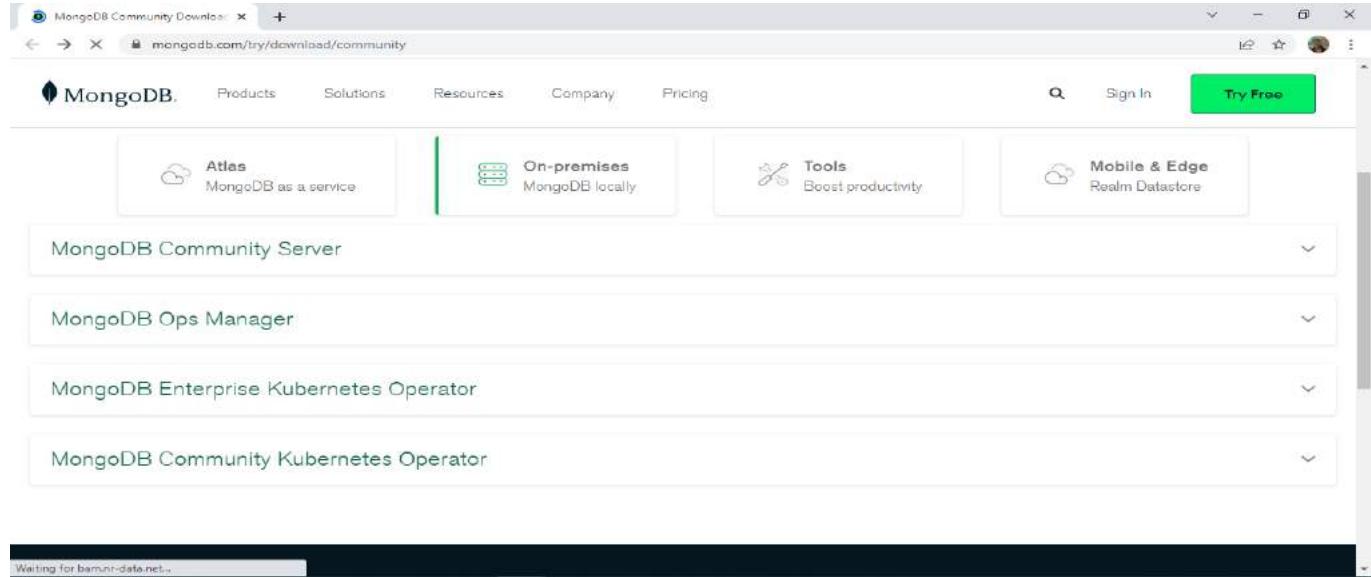
What is MongoDB?

MongoDB is a document-oriented **NoSQL database used for high volume data storage**. Instead of using tables and rows as in the traditional relational databases, MongoDB makes use of collections and documents.

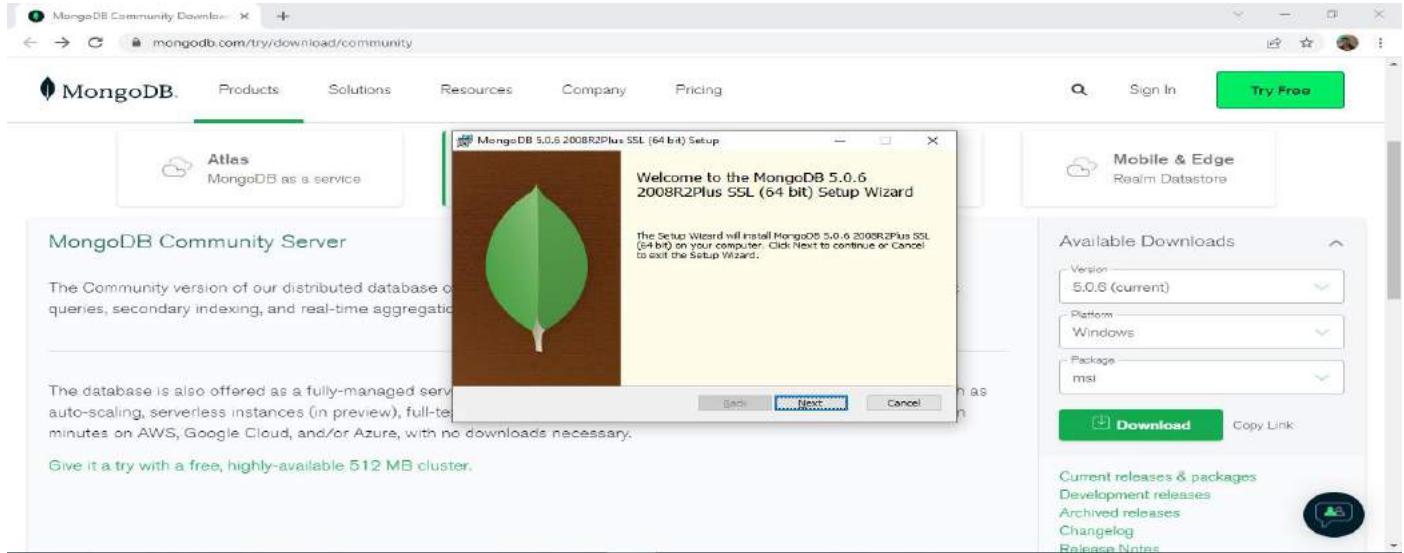
Documents consist of key-value pairs which are the basic unit of data in MongoDB. Collections contain sets of documents and function which is the equivalent of relational database tables. MongoDB is a database which came into light around the mid-2000s.

Download & Install MongoDB on Windows

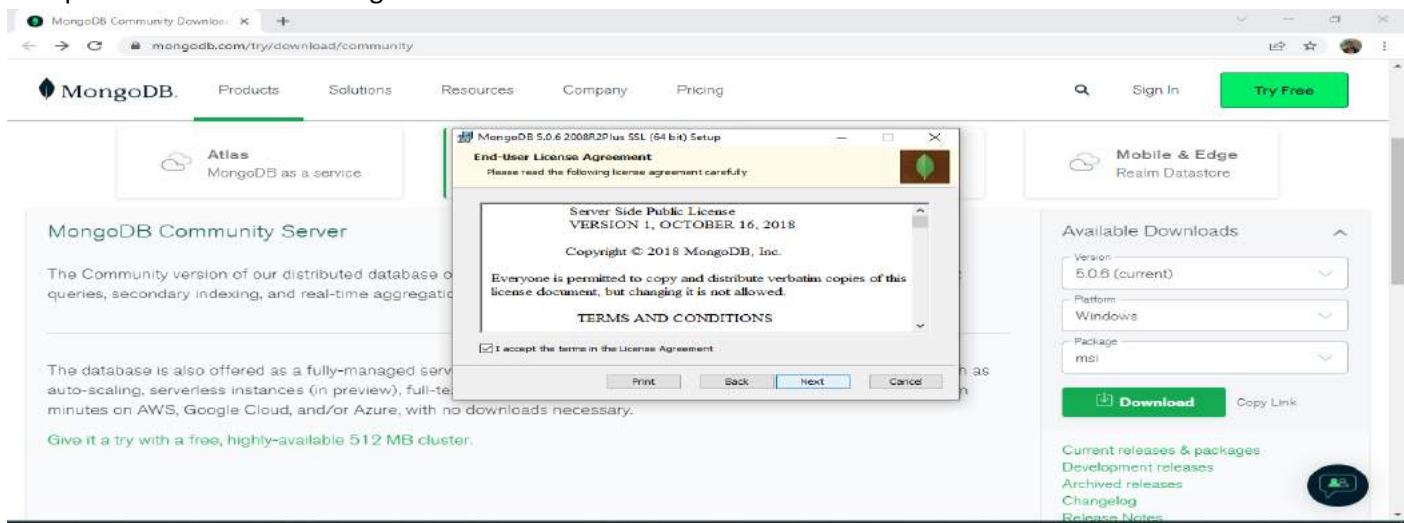
- 1) Go to link (<https://www.mongodb.com/try/download/community>) and Download MongoDB Community Server. We will install the 64-bit version for Windows.



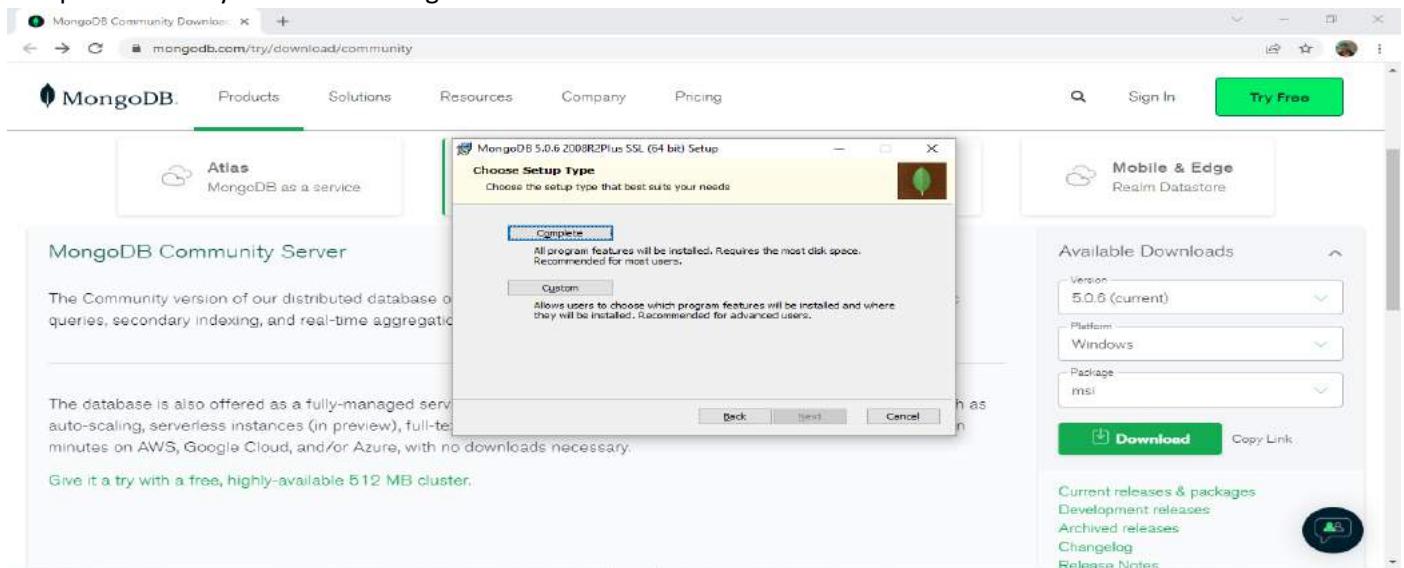
- 2) Once download is complete open the msi file. Click Next in the start-up screen



3) Accept the End-User License Agreement and Click Next



4) Click on the "complete" button to install all of the components. The custom option can be used to install selective components or if you want to change the location of the installation.



5) Select “Run service as Network Service user”. Make a note of the data directory, we’ll need this later and Click Next.

The screenshot shows the MongoDB Community Server download page. A 'Service Configuration' dialog box is open in the center, prompting for service setup details like account name, service name, and data/log directory paths. To the right, the 'Available Downloads' section is visible, showing 'Version: 5.0.6 (current)', 'Platform: Windows', and 'Package: msi'. A large green 'Download' button is prominent.

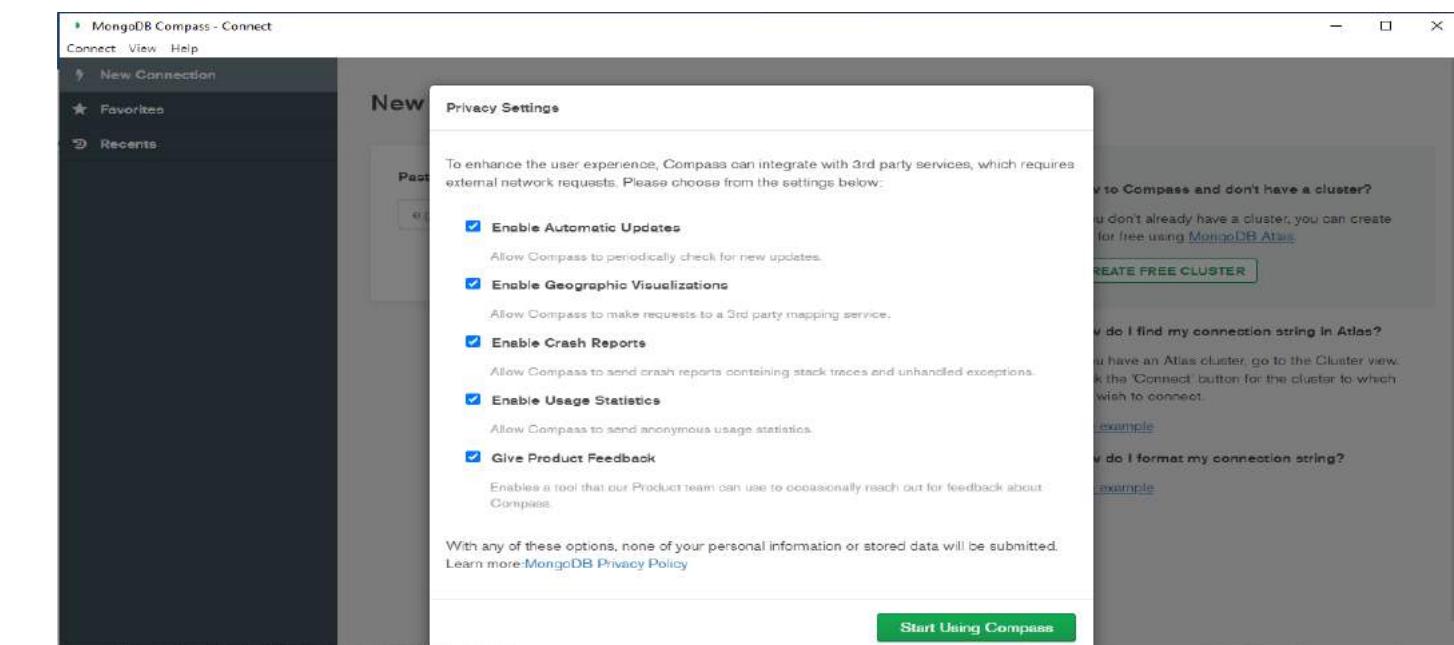
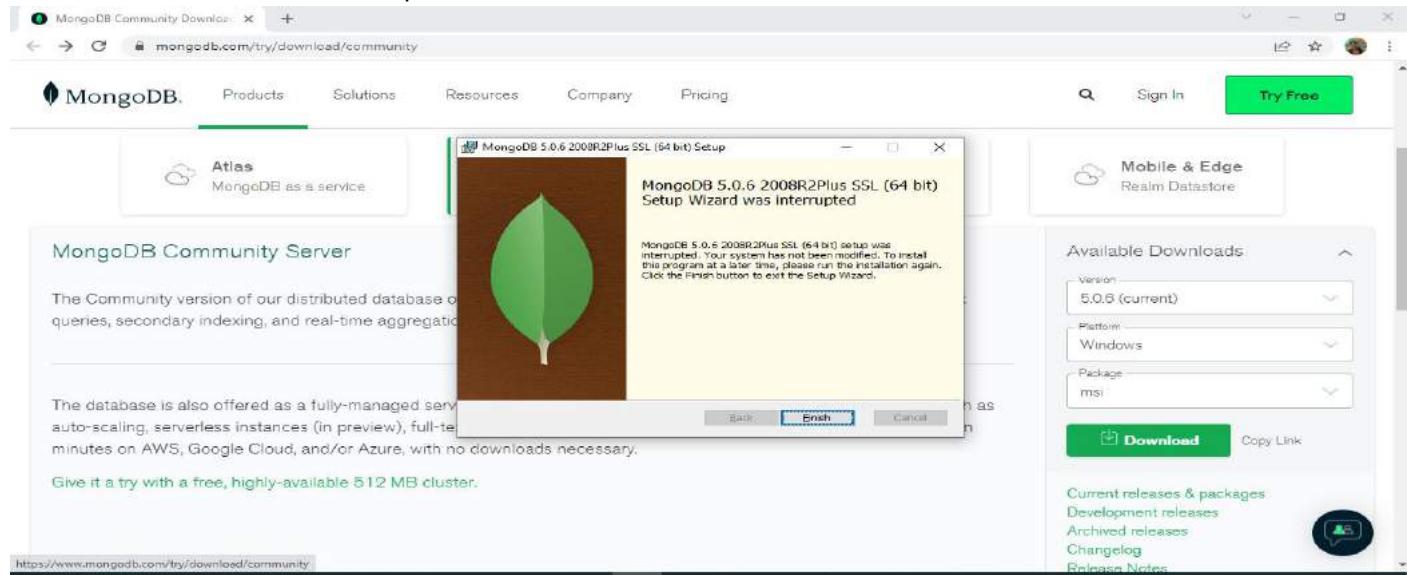
6) Click on the Install button to start the installation.

The screenshot shows the MongoDB Community Server download page. A 'Install MongoDB Compass' dialog box is open, indicating the installer will automatically download and install the latest version of MongoDB Compass. The 'Available Downloads' section on the right shows 'Version: 5.0.6 (current)', 'Platform: Windows', and 'Package: msi'.

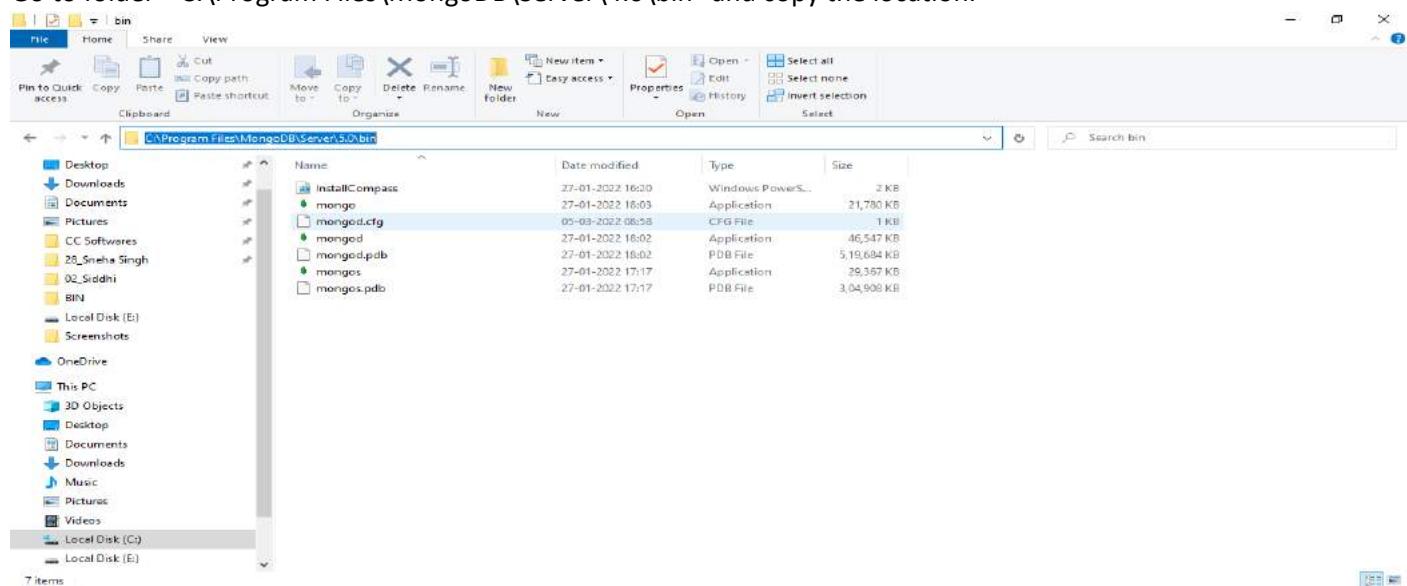
The screenshot shows the MongoDB Community Server download page. A 'Ready to install MongoDB 5.0.6 2008R2Plus SSL (64 bit)' dialog box is open, prompting to begin the installation. The 'Available Downloads' section on the right shows 'Version: 5.0.6 (current)', 'Platform: Windows', and 'Package: msi'.

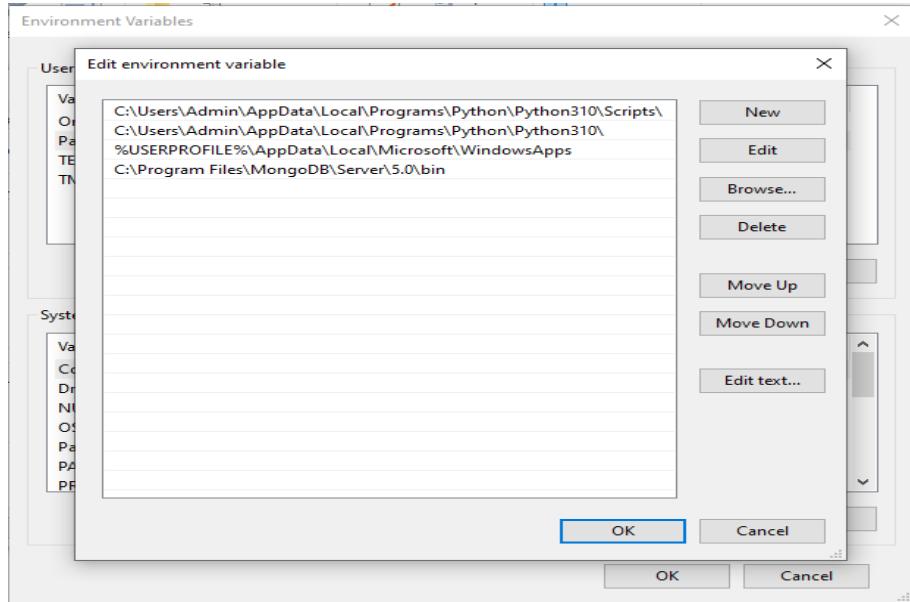
The screenshot shows the MongoDB Community Server download page. An 'Installing MongoDB 5.0.6 2008R2Plus SSL (64 bit)' progress dialog box is open, showing the status of the setup wizard. The 'Available Downloads' section on the right shows 'Version: 5.0.6 (current)', 'Platform: Windows', and 'Package: msi'.

7) Click on the Finish button to complete the installation



8) Go to folder " C:\Program Files\MongoDB\Server\4.0\bin" and copy the location.





- 9) Open the command prompt and run the cd command to change directory and run the command "mongod" to start the server and run the command "mongo" to work mongoDB.

```
C:\> Command Prompt
Microsoft Windows [Version 10.0.18363.657]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\> cd C:\Program Files\MongoDB\Server\5.0\bin
C:\Program Files\MongoDB\Server\5.0\bin> mongod
{"t": {"$date": "2022-03-05T09:03:29.600+05:30"}, "s": "I", "c": "CONTROL", "id": 23285, "ctx": "-", "msg": "Automatically disabling TLS 1.0, to force-enable TLS 1.0 specify --sslDisabledProtocols 'none'"}
{"t": {"$date": "2022-03-05T09:03:29.601+05:30"}, "s": "I", "c": "NETWORK", "id": 4915701, "ctx": "main", "msg": "Initialized wire specification", "attr": {"spec": {"incomingExternalClient": {"minWireVersion": 0, "maxWireVersion": 13}, "incomingInternalClient": {"minWireVersion": 0, "maxWireVersion": 13}, "outgoing": {"minWireVersion": 0, "maxWireVersion": 13}}, "isInternalClient": true}}
{"t": {"$date": "2022-03-05T09:03:29.607+05:30"}, "s": "W", "c": "ASIO", "id": 22601, "ctx": "main", "msg": "No TransportLayer configured during NetworkInterface startup"}
{"t": {"$date": "2022-03-05T09:03:29.608+05:30"}, "s": "I", "c": "NETWORK", "id": 4648602, "ctx": "main", "msg": "Implicit TCP FastOpen in use."}
{"t": {"$date": "2022-03-05T09:03:29.611+05:30"}, "s": "W", "c": "ASIO", "id": 22601, "ctx": "main", "msg": "No TransportLayer configured during NetworkInterface startup"}
{"t": {"$date": "2022-03-05T09:03:29.611+05:30"}, "s": "I", "c": "REPL", "id": 5123008, "ctx": "main", "msg": "Successfully registered PrimaryOnlyService", "attr": {"service": "TenantMigrationDonorService", "ns": "config.tenantMigrationDonors"}}
{"t": {"$date": "2022-03-05T09:03:29.612+05:30"}, "s": "I", "c": "REPL", "id": 5123008, "ctx": "main", "msg": "Successfully registered PrimaryOnlyService", "attr": {"service": "TenantMigrationRecipientsService", "ns": "config.tenantMigrationRecipients"}}
{"t": {"$date": "2022-03-05T09:03:29.615+05:30"}, "s": "I", "c": "CONTROL", "id": 5045603, "ctx": "main", "msg": "Multi threading initialized"}
{"t": {"$date": "2022-03-05T09:03:29.618+05:30"}, "s": "I", "c": "CONTROL", "id": 4615611, "ctx": "initandlisten", "msg": "MongoDB starting", "attr": {"pid": 3324, "port": 27017, "dbPath": "C:/data/db/", "architecture": "64-bit", "host": "PC-92"}}
{"t": {"$date": "2022-03-05T09:03:29.619+05:30"}, "s": "I", "c": "CONTROL", "id": 23398, "ctx": "initandlisten", "msg": "Target operating system minimum version", "attr": {"targetMinOS": "Windows 7/Windows Server 2008 R2"}}
{"t": {"$date": "2022-03-05T09:03:29.619+05:30"}, "s": "", "c": "CONTROL", "id": 23403, "ctx": "initandlisten", "msg": "Build Info", "attr": {"buildInfo": {"version": "5.0.6", "gitVersion": "212a8dbb47f07427dae1949c75baec1d81d9259", "modules": [], "allocator": "tcmalloc", "environment": {"distmod": "windows", "distarch": "x86_64", "target_arch": "x86_64"}}}
{"t": {"$date": "2022-03-05T09:03:29.619+05:30"}, "s": "I", "c": "CONTROL", "id": 51765, "ctx": "initandlisten", "msg": "Operating System", "attr": {"os": {"name": "Microsoft Windows 10", "version": "10.0 (build 18363)"}}, "attr": {"options": {}}}
{"t": {"$date": "2022-03-05T09:03:29.619+05:30"}, "s": "I", "c": "CONTROL", "id": 21951, "ctx": "initandlisten", "msg": "Options set by command line", "attr": {"options": {}}}
{"t": {"$date": "2022-03-05T09:03:29.622+05:30"}, "s": "E", "c": "CONTROL", "id": 20557, "ctx": "initandlisten", "msg": "DBException in initAndListen, terminating", "attr": {"error": "NonExistentPath: Data directory C:\\data\\db\\ not found. Create the missing directory or specify another path using (1) the --dbpath command line option, or (2) by adding the 'storage.dbPath' option in the configuration file."}}
{"t": {"$date": "2022-03-05T09:03:29.623+05:30"}, "s": "I", "c": "REPL", "id": 4784900, "ctx": "initandlisten", "msg": "Stepping down the ReplicationCoordinator for shutdown", "attr": {"waitTimeMillis": 15000}}
{"t": {"$date": "2022-03-05T09:03:29.625+05:30"}, "s": "", "c": "COMMAND", "id": 4784901, "ctx": "initandlisten", "msg": "Shutting down the MirrorMaestro"}
{"t": {"$date": "2022-03-05T09:03:29.628+05:30"}, "s": "I", "c": "SHARDING", "id": 4784902, "ctx": "initandlisten", "msg": "Shutting down the WaitForMajorityService"}
{"t": {"$date": "2022-03-05T09:03:29.628+05:30"}, "s": "I", "c": "NETWORK", "id": 20562, "ctx": "initandlisten", "msg": "Shutdown: going to close listening sockets"}
{"t": {"$date": "2022-03-05T09:03:29.630+05:30"}, "s": "I", "c": "NETWORK", "id": 4784905, "ctx": "initandlisten", "msg": "Shutting down the global connection pool"}
{"t": {"$date": "2022-03-05T09:03:29.630+05:30"}, "s": "I", "c": "CONTROL", "id": 4784906, "ctx": "initandlisten", "msg": "Shutting down the FlowControlTicketholder"}
{"t": {"$date": "2022-03-05T09:03:29.631+05:30"}, "s": "I", "c": "ASIO", "id": 22582, "ctx": "MigrationUtil-TaskExecutor", "msg": "Killing all outstanding egress activity"}
```

Aim: B- Performing CRUD Operations for unstructured data

For storing data in a MongoDB, you need to create a database first. It will allow you to systematically organize your data so that it can be retrieved as per requirement. If you wish to delete a database, MongoDB also allows you to delete that.

What is CRUD in MongoDB?

CRUD operations describe the conventions of a user-interface that let users view, search, and modify parts of the database.

MongoDB documents are modified by connecting to a server, querying the proper documents, and then changing the setting properties before sending the data back to the database to be updated.

CRUD is data-oriented, and it's standardized according to HTTP action verbs.

When it comes to the individual CRUD operations:

1. The Create operation is used to insert new documents in the MongoDB database.
2. The Read operation is used to query a document in the database.
3. The Update operation is used to modify existing documents in the database.
4. The Delete operation is used to remove documents in the database.

Creating a Database in MongoDB

MongoDB has no "create" command for creating a database. Also, it is essential to note that MongoDB does not provide any specific command for creating a database. This might seem a bit agitated if you are new to this subject and database tool or in case you have used that conventional SQL as your database where you are required to create a new database, which will contain table and, you will then have to use the INSERT INTO TABLE to insert values manually within your table.

In this MongoDB tool, you need not have to produce, or it is optional to create a database manually. This is because MongoDB has the feature of automatically creating it for the first time for you, once you save your value in that collection. So, explicitly, you do not need to mention or put a command to create a database; instead it will be created automatically once the collection is filled with values.

The "use" Command for Creating Database in MongoDB

You can make use of the "use" command followed by the database_name for creating a database. This command will tell the MongoDB client to create a database by this name if there is no database exists by this name. Otherwise, this command will return the existing database that has the name.

Show list of databases

First start the mongo Db using command **mongo**

1) show dbs;

```
Command Prompt - mongo
For installation instructions, see
https://docs.mongodb.com/mongodb-shell/install/
=====
Welcome to the MongoDB shell.
For interactive help, type "help".
For more comprehensive documentation, see
    https://docs.mongodb.com/
Questions? Try the MongoDB Developer Community Forums
    https://community.mongodb.com
---
The server generated these startup warnings when booting:
2022-03-05T08:58:27.721+05:30: Access control is not enabled for the database. Read and write access to data and
configuration is unrestricted
---
Enable MongoDB's free cloud-based monitoring service, which will then receive and display
metrics about your deployment (disk utilization, CPU, operation statistics, etc).
The monitoring data will be available on a MongoDB website with a unique URL accessible to you
and anyone you share the URL with. MongoDB may use this information to make product
improvements and to suggest MongoDB products and deployment options to you.
To enable free monitoring, run the following command: db.enableFreeMonitoring()
To permanently disable this reminder, run the following command: db.disableFreeMonitoring()
---
> show dbs;
admin      0.000GB
config     0.000GB
local      0.000GB
>
```

2) use RetailBikeDb

```
Command Prompt - mongo
> show dbs;
admin      0.000GB
config     0.000GB
local      0.000GB
> use RetailBikeDb
switched to db RetailBikeDb
> -
```

3) Creating and showing collections in MongoDB

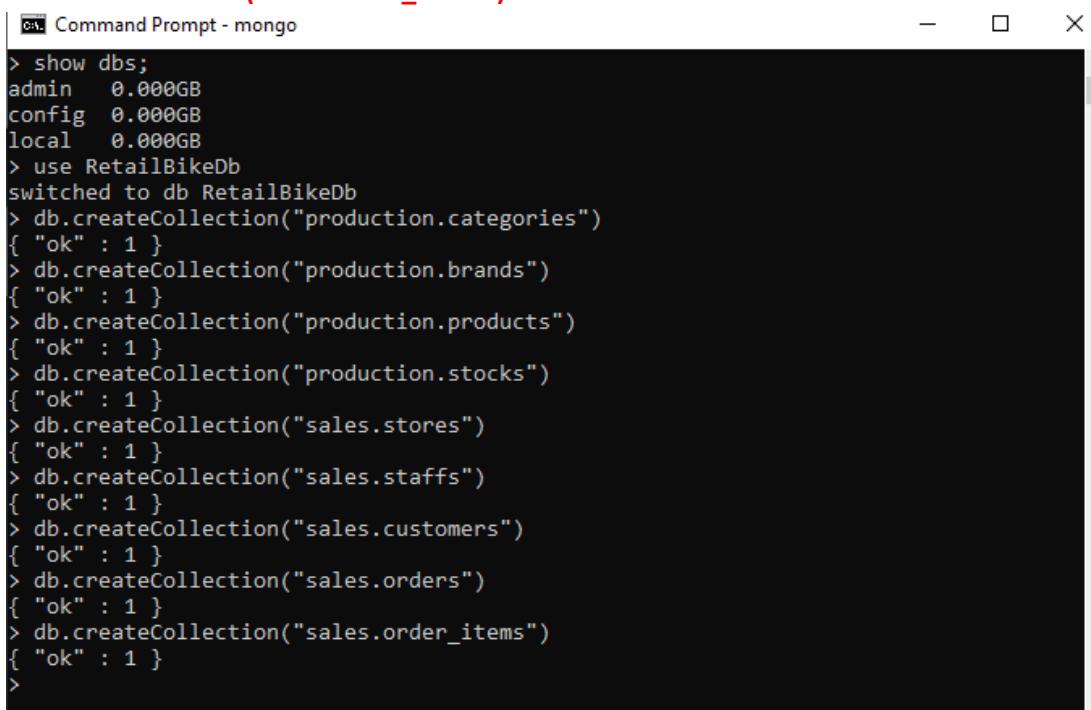
Collections are like that of tables of RDBMS and are capable enough to store documents of diverse or dissimilar types.

Creation and removal of collections in MongoDB can be done in specific ways. In this chapter, you will learn about the creation of collections in a database created using MongoDB.

Creation of collection can be done using

Syntax: db.createCollection(name)

```
db.createCollection("production.categories")
db.createCollection("production.brands")
db.createCollection("production.products")
db.createCollection("production.stocks")
db.createCollection("sales.stores")
db.createCollection("sales.staffs")
db.createCollection("sales.customers")
db.createCollection("sales.orders")
db.createCollection("sales.order_items")
```



The screenshot shows a terminal window titled "Command Prompt - mongo". The window contains a list of MongoDB commands used to create collections in a database named "RetailBikeDb". The commands are:

```
> show dbs;
admin 0.000GB
config 0.000GB
local 0.000GB
> use RetailBikeDb
switched to db RetailBikeDb
> db.createCollection("production.categories")
{ "ok" : 1 }
> db.createCollection("production.brands")
{ "ok" : 1 }
> db.createCollection("production.products")
{ "ok" : 1 }
> db.createCollection("production.stocks")
{ "ok" : 1 }
> db.createCollection("sales.stores")
{ "ok" : 1 }
> db.createCollection("sales.staffs")
{ "ok" : 1 }
> db.createCollection("sales.customers")
{ "ok" : 1 }
> db.createCollection("sales.orders")
{ "ok" : 1 }
> db.createCollection("sales.order_items")
{ "ok" : 1 }
>
```

4) We can show list of collection using "Show collections" commands in MongoDB.

show collections

```
> show collections
production.brands
production.categories
production.products
production.stocks
sales.customers
sales.order_items
sales.orders
sales.staffs
sales.stores
>
```

5) CRUD operation

CRUD operation is one of the essential concepts of a database system. Inserting data in the database comes under one of the CRUD operations. If you do not insert data in your database, you will not be able to continue with other activities within your document.

1) Create Operations

For MongoDB CRUD, if the specified collection doesn't exist, the create operation will create the collection when it's executed. Create operations in MongoDB target a single collection, not multiple collections. Insert operations in MongoDB are atomic on a single document level.

MongoDB provides two different create operations that you can use to insert documents into a collection:

1. **db.collection.insertOne()**
2. **db.collection.insertMany()**

The insertOne() method

As the namesake, insertOne() allows you to insert one document into the collection. Example -

```
db.movie.insertOne({ _id: 2, writername: "Stan Lee", name: "Aquaman" })
```

The insertMany() method

It's possible to insert multiple items at one time by calling the insertMany() method on the desired collection. In this case, we pass multiple items into our chosen collection (RetailDB) and separate them by commas. Within the parentheses, we use brackets to indicate that we are passing in a list of multiple entries. This is commonly referred to as a nested method.

I have taken example of Retail Bike store database to demonstrate insert crud operations.

1. **production.categories**
db.production.categories.insertMany(
[
{
"category_id": 1,
"category_name": "Road Bike"
},
{
"category_id": 2,
"category_name": "Mountain Bike"
},
{
"category_id": 3,
"category_name": "Hybrid Bike"
},
{
"category_id": 4,
"category_name": "Folding Bike"
},
{
"category_id": 5,
"category_name": "Touring Bike"
},
{
"category_id": 6,
"category_name": "Cruiser Bike"
},
{
"category_id": 7,
"category_name": "Women Bike"
}
]

```
Command Prompt - mongo
sales.stores
> db.production.categories.insertMany(
... [
... {
... "category_id": 1,
... "category_name": "Road Bike"
... },
... {
... "category_id": 2,
... "category_name": "Mountain Bike"
... },
... {
... "category_id": 3,
... "category_name": "Hybrid Bike"
... },
... {
... "category_id": 4,
... "category_name": "Folding Bike"
... },
... {
... "category_id": 5,
... "category_name": "Touring Bike"
... },
... {
... "category_id": 6,
... "category_name": "Cruiser Bike"
... },
... {
... "category_id": 7,
... "category_name": "Women Bike"
... }
... ]
... )
{
    "acknowledged" : true,
    "insertedIds" : [
        ObjectId("6222df12b5e511acae1d7891"),
        ObjectId("6222df12b5e511acae1d7892"),
        ObjectId("6222df12b5e511acae1d7893"),
        ObjectId("6222df12b5e511acae1d7894"),
        ObjectId("6222df12b5e511acae1d7895"),
        ObjectId("6222df12b5e511acae1d7896"),
        ObjectId("6222df12b5e511acae1d7897")
    ]
}
```

2. production.products

```
db.production.products.insertMany(
[
{
"product_id": 1,
"product_name": "Honda Superfast",
"brand_id": 1,
"category_id": 1,
"model_year": 1994,
"list_price": 25000
},
{
"product_id": 2,
"product_name": "6KU Bikes",
"brand_id": 2,
"category_id": 4,
"model_year": 2000,
"list_price": 30000
},
{
"product_id": 3,
"product_name": "Bianchi",
"brand_id": 3,
"category_id": 2,
"model_year": 2002,
"list_price": 30000
},
{
}
```

```
"product_id": 4,  
"product_name": "BMC Hybrid Bike",  
"brand_id": 4,  
"category_id": 3,  
"model_year": 2009,  
  
"list_price": 45000  
},  
{  
"product_id": 5,  
"product_name": "Huffy Women Bike",  
"brand_id": 5,  
"category_id": 7,  
"model_year": 2019,  
"list_price": 50000  
}  
]  
)
```

```
 Command Prompt - mongo  
... "product_name": "6KU Bikes",  
... "brand_id": 2,  
... "category_id": 4,  
... "model_year": 2000,  
... "list_price": 30000  
},  
{  
... "product_id": 3,  
... "product_name": "Bianchi",  
... "brand_id": 3,  
... "category_id": 2,  
... "model_year": 2002,  
... "list_price": 30000  
},  
{  
... "product_id": 4,  
... "product_name": "BMC Hybrid Bike",  
... "brand_id": 4,  
... "category_id": 3,  
... "model_year": 2009,  
... "list_price": 45000  
},  
{  
... "product_id": 5,  
... "product_name": "Huffy Women Bike",  
... "brand_id": 5,  
... "category_id": 7,  
... "model_year": 2019,  
... "list_price": 50000  
}  
]  
)  
{  
    "acknowledged" : true,  
    "insertedIds" : [  
        ObjectId("6222df37b5e511acae1d7898"),  
        ObjectId("6222df37b5e511acae1d7899"),  
        ObjectId("6222df37b5e511acae1d789a"),  
        ObjectId("6222df37b5e511acae1d789b"),  
        ObjectId("6222df37b5e511acae1d789c")  
    ]  
}
```

3. production.stocks

```
db.production.stocks.insertMany(  
[  
{  
    "store_id": 1,  
    "product_id": 1,  
    "quantity": 15  
},  
{  
    "store_id": 2,  
    "product_id": 4,  
    "quantity": 20  
},  
{  
    "store_id": 3,  
    "product_id": 5,  
    "quantity": 30  
},  
{  
    "store_id": 1,  
    "product_id": 5,  
    "quantity": 100  
},  
{  
    "store_id": 2,  
    "product_id": 2,  
    "quantity": 4  
},  
{  
    "store_id": 3,  
    "product_id": 3,  
    "quantity": 9  
}  
]  
)
```

```
... {  
...     "store_id": 1,  
...     "product_id": 1,  
...     "quantity": 15  
... },  
... {  
...     "store_id": 2,  
...     "product_id": 4,  
...     "quantity": 20  
... },  
... {  
...     "store_id": 3,  
...     "product_id": 5,  
...     "quantity": 30  
... },  
... {  
...     "store_id": 1,  
...     "product_id": 5,  
...     "quantity": 100  
... },  
... {  
...     "store_id": 2,  
...     "product_id": 2,  
...     "quantity": 4  
... },  
... {  
...     "store_id": 3,  
...     "product_id": 3,  
...     "quantity": 9  
}  
]  
)  
{  
    "acknowledged" : true,  
    "insertedIds" : [  
        ObjectId("6222dfd4b5e511acae1d789d"),  
        ObjectId("6222dfd4b5e511acae1d789e"),  
        ObjectId("6222dfd4b5e511acae1d789f"),  
        ObjectId("6222dfd4b5e511acae1d78a0"),  
        ObjectId("6222dfd4b5e511acae1d78a1"),  
        ObjectId("6222dfd4b5e511acae1d78a2")  
    ]  
}
```

4. sales.customers

```
db.sales.customers.insertMany(  
[  
{  
"customer_id": "Cus001",  
"first_name": "Jay",  
"last_name": "Mehta",  
"phone": 1234567890,  
"email": "jay@gmail.com",  
"street": "T.P Road",  
"city": "Mumbai",  
"state": "Maharashtra",  
"zip_code": 400072  
},  
{  
"customer_id": "Cus002",  
"first_name": "Ruhি",  
"last_name": "Singh",  
"phone": 7894561230,  
"email": "ruhi@yahoo.com",  
"street": "M.G Chauk",  
"city": "Mumbai",  
"state": "Maharashtra",  
"zip_code": 400072  
},  
{  
"customer_id": "Cus003",  
"first_name": "Aria",  
"last_name": "Josh",  
"phone": 1245789630,  
"email": "aria.josh@gmail.com",  
"street": "JVM",  
"city": "Gandhi Nagar",  
  
"state": "Gujrat",  
"zip_code": 401235  
},  
{  
"customer_id": "Cus004",  
"first_name": "Mahi",  
"last_name": "Kaur",  
"phone": 4567890123,  
"email": "kaur.mahi@hotmail.com",  
"street": "J.V.L.R",  
"city": "Mumbai",  
"state": "Maharashtra",  
"zip_code": 400072  
},  
{  
"customer_id": "Cus005",  
"first_name": "Aditya",  
"last_name": "Yadav",  
"phone": 9638527410,  
"email": "aditya@gmail.com",  
"street": "Koliwada",  
}]
```

```
        "city": "Pune",
        "state": "Maharashtra",
        "zip_code": 300075
    }
]
)
```

The screenshot shows a terminal window titled "Select Command Prompt - mongo". The output displays the results of an insertMany operation into a collection. The data inserted consists of five documents, each representing a customer with fields like last_name, first_name, email, phone, street, city, state, and zip_code. Each document also includes an "_id" field represented as an ObjectId. The final part of the output shows the response from the database, indicating successful insertion with acknowledged=true and a list of insertedIds.

```
... "last_name": "Josh",
... "phone": 1245789630,
... "email": "aria.josh@gmail.com",
... "street": "JVM",
... "city": "Gandhi Nagar",
...
... "state": "Gujrat",
... "zip_code": 401235
},
{
...
"customer_id": "Cus004",
"first_name": "Mahi",
"last_name": "Kaur",
"phone": 4567890123,
"email": "kaur.mahi@hotmail.com",
"street": "J.V.L.R",
"city": "Mumbai",
"state": "Maharashtra",
"zip_code": 400072
},
{
...
"customer_id": "Cus005",
"first_name": "Aditya",
"last_name": "Yadav",
"phone": 9638527410,
"email": "aditya@gmail.com",
"street": "Koliwada",
"city": "Pune",
"state": "Maharashtra",
"zip_code": 300075
}
]
)
{
    "acknowledged" : true,
    "insertedIds" : [
        ObjectId("6222e05db5e511acae1d78a3"),
        ObjectId("6222e05db5e511acae1d78a4"),
        ObjectId("6222e05db5e511acae1d78a5"),
        ObjectId("6222e05db5e511acae1d78a6"),
        ObjectId("6222e05db5e511acae1d78a7")
    ]
}
>
```

5. sales.order_items

```
db.sales.order_items.insertMany(
[
{
    "order_id": "ORD001",
    "product_id": 1,
    "quantity": 2,
    "list_price": 50000
},
{
    "order_id": "ORD002",
    "product_id": 2,
    "quantity": 3,
    "list_price": 90000
},
{
    "order_id": "ORD003",
    "product_id": 3,
```

```
        "quantity": 1,  
        "list_price": 30000  
    },  
    {  
        "order_id": "ORD004",  
        "product_id": 4,  
        "quantity": 8,  
        "list_price": 360000  
    },  
    {  
        "order_id": "ORD005",  
        "product_id": 5,  
        "quantity": 2,  
        "list_price": 100000  
    }  
  
]  
)
```

```
0: Command Prompt - mongo  
... {  
...     "order_id": "ORD001",  
...     "product_id": 1,  
...     "quantity": 2,  
...     "list_price": 50000  
... },  
... {  
...     "order_id": "ORD002",  
...     "product_id": 2,  
...     "quantity": 3,  
...     "list_price": 90000  
... },  
... {  
...     "order_id": "ORD003",  
...     "product_id": 3,  
...     "quantity": 1,  
...     "list_price": 30000  
... },  
... {  
...     "order_id": "ORD004",  
...     "product_id": 4,  
...     "quantity": 8,  
...     "list_price": 360000  
... },  
... {  
...     "order_id": "ORD005",  
...     "product_id": 5,  
...     "quantity": 2,  
...     "list_price": 100000  
... }  
... ]  
... )  
{  
    "acknowledged" : true,  
    "insertedIds" : [  
        ObjectId("6222e0bab5e511acae1d78a8"),  
        ObjectId("6222e0bab5e511acae1d78a9"),  
        ObjectId("6222e0bab5e511acae1d78aa"),  
        ObjectId("6222e0bab5e511acae1d78ab"),  
        ObjectId("6222e0bab5e511acae1d78ac")  
    ]  
}  
>
```

6. sales.orders

```
db.sales.orders.insertMany(  
[  
  {  
    "order_id": "ORD001",  
    "customer_id": "Cus001",  
    "order_status": "Completed",  
    "order_date": 43992,  
    "shipped_date": 43994,  
    "store_id": 1,  
    "staff_id": 1  
  },  
  {  
    "order_id": "ORD002",  
    "customer_id": "Cus002",  
    "order_status": "Completed",  
    "order_date": 44221,  
    "shipped_date": 44227,  
    "store_id": 2,  
    "staff_id": 2  
  },  
  {  
    "order_id": "ORD003",  
    "customer_id": "Cus003",  
    "order_status": "Completed",  
    "order_date": 44306,  
    "shipped_date": 44314,  
    "store_id": 2,  
    "staff_id": 2  
  },  
  {  
    "order_id": "ORD004",  
    "customer_id": "Cus004",  
    "order_status": "Pending",  
    "order_date": 44367,  
    "shipped_date": 44377,  
    "store_id": 3,  
    "staff_id": 3  
  },  
  {  
    "order_id": "ORD005",  
    "customer_id": "Cus005",  
    "order_status": "Pending",  
    "order_date": 44367,  
    "shipped_date": 44377,  
    "store_id": 1,  
    "staff_id": 1  
  }  
]
```

```
... Command Prompt - mongo
...
... "store_id": 2,
... "staff_id": 2
... },
...
... {
...   "order_id": "ORD003",
...   "customer_id": "Cus003",
...   "order_status": "Completed",
...   "order_date": 44306,
...   "shipped_date": 44314,
...   "store_id": 2,
...   "staff_id": 2
...
... },
...
... {
...   "order_id": "ORD004",
...   "customer_id": "Cus004",
...   "order_status": "Pending",
...   "order_date": 44367,
...   "shipped_date": 44377,
...   "store_id": 3,
...   "staff_id": 3
... },
...
... {
...   "order_id": "ORD005",
...   "customer_id": "Cus005",
...   "order_status": "Pending",
...   "order_date": 44367,
...   "shipped_date": 44377,
...   "store_id": 1,
...   "staff_id": 1
... }
...
...
...
{
  "acknowledged" : true,
  "insertedIds" : [
    ObjectId("6222e0f7b5e511acae1d78ad"),
    ObjectId("6222e0f7b5e511acae1d78ae"),
    ObjectId("6222e0f7b5e511acae1d78af"),
    ObjectId("6222e0f7b5e511acae1d78b0"),
    ObjectId("6222e0f7b5e511acae1d78b1")
  ]
}
> -
```

7. sales.staffs

```
db.sales.staffs.insertMany(  
[  
 {  
 "staff_id": 1,  
 "first_name": "Pushpa",  
 "last_name": "Yadav",  
 "email": "pushpa@gmail.com",  
 "phone": 9999999999,  
 "active": "Yes",  
 "store_id": 1,  
 "manager_id": 1  
 },  
 {  
 "staff_id": 2,  
 "first_name": "Sadiksha",  
 "last_name": "Singh",  
 "email": "sadiksha@gmail.com",  
 "phone": 8888888888,  
 "active": "Yes",  
 "store_id": 2,  
 "manager_id": 1  
 },  
 {  
 "staff_id": 3,  
 "first_name": "Priya",  
 "last_name": "Kumar",  
 "email": "priyakumar@gmail.com",  
 "phone": 9898989898,  
 "active": "Yes",  
 "store_id": 3,  
 "manager_id": 2  
 }]
```

```
"last_name": "Nadar",
"email": "priya@gmail.com",
"phone": 7777777777,
"active": "Yes",
"store_id": 3,
"manager_id": 1
```

```
}
```

```
]
```

```
)
```

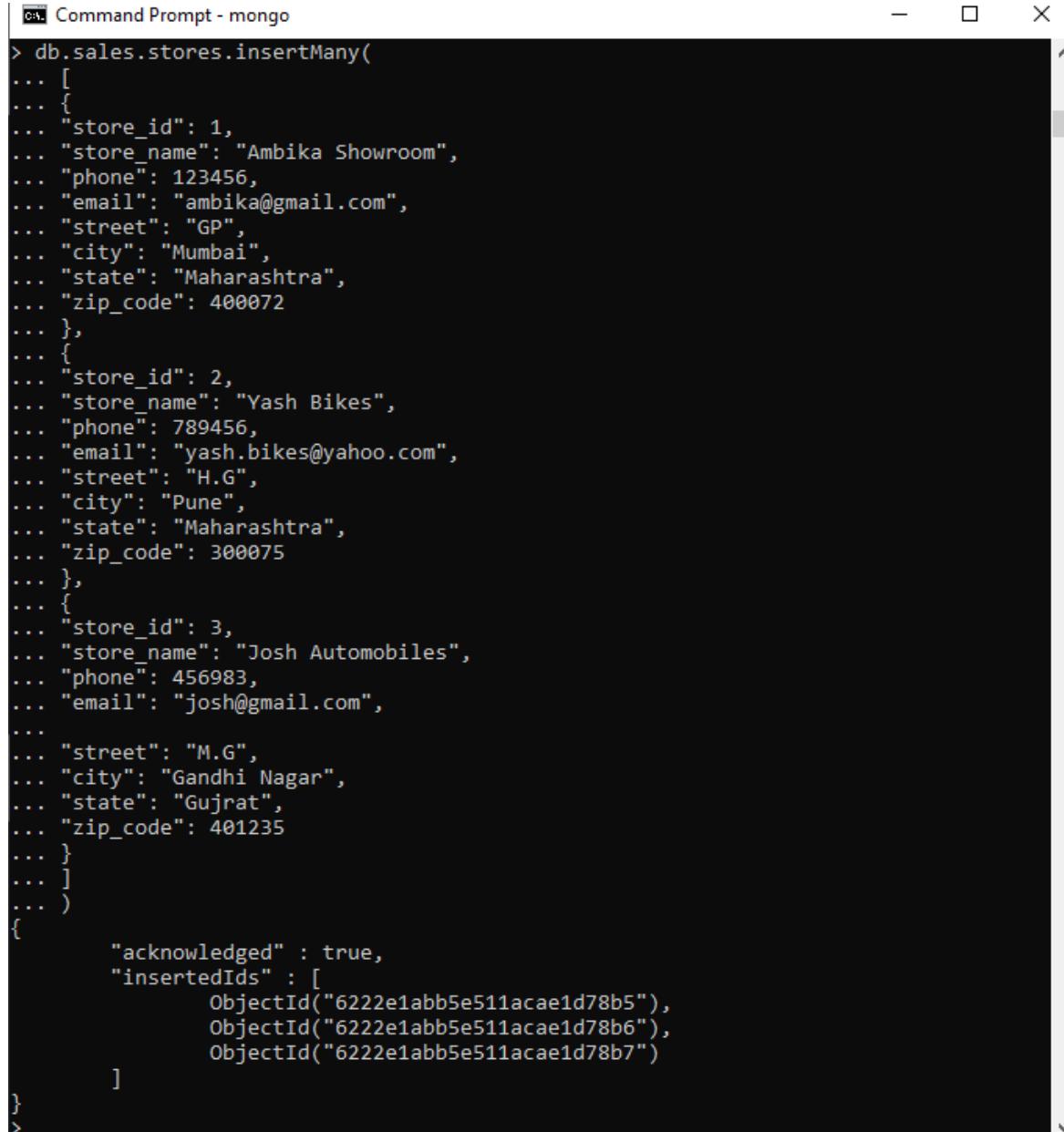
```
| C:\ Command Prompt - mongo
```

```
> db.sales.staffs.insertMany(
... [
... {
... "staff_id": 1,
... "first_name": "Pushpa",
... "last_name": "Yadav",
... "email": "pushpa@gmail.com",
... "phone": 9999999999,
... "active": "Yes",
... "store_id": 1,
... "manager_id": 1
... },
... {
... "staff_id": 2,
... "first_name": "Sadiksha",
... "last_name": "Singh",
... "email": "sadiksha@gmail.com",
... "phone": 8888888888,
... "active": "Yes",
... "store_id": 2,
... "manager_id": 1
... },
... {
... "staff_id": 3,
... "first_name": "Priya",
... "last_name": "Nadar",
... "email": "priya@gmail.com",
... "phone": 7777777777,
... "active": "Yes",
... "store_id": 3,
... "manager_id": 1
...
... }
... ]
... )
{
    "acknowledged" : true,
    "insertedIds" : [
        ObjectId("6222e161b5e511acae1d78b2"),
        ObjectId("6222e161b5e511acae1d78b3"),
        ObjectId("6222e161b5e511acae1d78b4")
    ]
}
>
```

8. sales.stores

```
db.sales.stores.insertMany(
[
{
"store_id": 1,
"store_name": "Ambika Showroom",
"phone": 123456,
"email": "ambika@gmail.com",
"street": "GP",
"city": "Mumbai",
"state": "Maharashtra",
"zip_code": 400072
},
{
```

```
"store_id": 2,  
"store_name": "Yash Bikes",  
"phone": 789456,  
"email": "yash.bikes@yahoo.com",  
"street": "H.G",  
"city": "Pune",  
"state": "Maharashtra",  
"zip_code": 300075  
},  
{  
"store_id": 3,  
"store_name": "Josh Automobiles",  
"phone": 456983,  
"email": "josh@gmail.com",  
  
"street": "M.G",  
"city": "Gandhi Nagar",  
"state": "Gujrat",  
"zip_code": 401235  
}  
]  
)
```



```
Command Prompt - mongo  
> db.sales.stores.insertMany(  
... [  
... {  
... "store_id": 1,  
... "store_name": "Ambika Showroom",  
... "phone": 123456,  
... "email": "ambika@gmail.com",  
... "street": "GP",  
... "city": "Mumbai",  
... "state": "Maharashtra",  
... "zip_code": 400072  
},  
... {  
... "store_id": 2,  
... "store_name": "Yash Bikes",  
... "phone": 789456,  
... "email": "yash.bikes@yahoo.com",  
... "street": "H.G",  
... "city": "Pune",  
... "state": "Maharashtra",  
... "zip_code": 300075  
},  
... {  
... "store_id": 3,  
... "store_name": "Josh Automobiles",  
... "phone": 456983,  
... "email": "josh@gmail.com",  
  
... "street": "M.G",  
... "city": "Gandhi Nagar",  
... "state": "Gujrat",  
... "zip_code": 401235  
}  
]  
)  
{  
    "acknowledged" : true,  
    "insertedIds" : [  
        ObjectId("6222e1abb5e511acae1d78b5"),  
        ObjectId("6222e1abb5e511acae1d78b6"),  
        ObjectId("6222e1abb5e511acae1d78b7")  
    ]  
}
```

9. production.brands

```
db.production.brands.insertMany(  
[  
{  
"brand_id": 1,  
"brand_name": "Honda"  
},  
{  
"brand_id": 2,  
"brand_name": "6KU Bikes"  
},  
{  
"brand_id": 3,  
"brand_name": "Bianchi"  
},  
{  
"brand_id": 4,  
"brand_name": "BMC"  
},  
{  
"brand_id": 5,  
"brand_name": "Huffy"  
}  
]
```

```
)
```

The screenshot shows a Command Prompt window titled "Command Prompt - mongo". The window displays the execution of a MongoDB command to insert multiple documents into the "brands" collection in the "production" database. The command is as follows:

```
{  
    "acknowledged" : true,  
    "insertedIds" : [  
        ObjectId("6222e1abb5e511acae1d78b5"),  
        ObjectId("6222e1abb5e511acae1d78b6"),  
        ObjectId("6222e1abb5e511acae1d78b7")  
    ]  
}  
> db.production.brands.insertMany(  
... [  
... {  
... "brand_id": 1,  
... "brand_name": "Honda"  
... },  
... {  
... "brand_id": 2,  
... "brand_name": "6KU Bikes"  
... },  
... {  
... "brand_id": 3,  
... "brand_name": "Bianchi"  
... },  
... {  
... "brand_id": 4,  
... "brand_name": "BMC"  
... },  
... {  
... "brand_id": 5,  
... "brand_name": "Huffy"  
... }  
... ]  
... )  
{  
    "acknowledged" : true,  
    "insertedIds" : [  
        ObjectId("6222e1dc5e511acae1d78b8"),  
        ObjectId("6222e1dc5e511acae1d78b9"),  
        ObjectId("6222e1dc5e511acae1d78ba"),  
        ObjectId("6222e1dc5e511acae1d78bb"),  
        ObjectId("6222e1dc5e511acae1d78bc")  
    ]  
}
```

2) Read Operations

The read operations allow you to supply special query filters and criteria that let you specify which documents you want. The MongoDB documentation contains more information on the available query filters. Query modifiers may also be used to change how many results are returned.

MongoDB has two methods of reading documents from a collection:

- **db.collection.find()**
- **db.collection.findOne()**

find()

In order to get all the documents from a collection, we can simply use the find() method on our chosen collection. Executing just the find() method with no arguments will return all records currently in the collection.

Example - db.production.brands.find()

Here we can see that every record has an assigned “ObjectId” mapped to the “_id” key.

```
> db.production.brands.find()
{ "_id" : ObjectId("6222e1dcb5e511acae1d78b8"), "brand_id" : 1, "brand_name" : "Honda" }
{ "_id" : ObjectId("6222e1dcb5e511acae1d78b9"), "brand_id" : 2, "brand_name" : "6KU Bikes" }
{ "_id" : ObjectId("6222e1dcb5e511acae1d78ba"), "brand_id" : 3, "brand_name" : "Bianchi" }
{ "_id" : ObjectId("6222e1dcb5e511acae1d78bb"), "brand_id" : 4, "brand_name" : "BMC" }
{ "_id" : ObjectId("6222e1dcb5e511acae1d78bc"), "brand_id" : 5, "brand_name" : "Huffy" }
>
```

If you want to get more specific with a read operation and find a desired subsection of the records, you can use the previously mentioned filtering criteria to choose what results should be returned. One of the most common ways of filtering the results is to search by value.

Example: db.production.brands.find({"brand_name" : "Honda"})

```
> db.production.brands.find({"brand_name" : "Honda"})
{ "_id" : ObjectId("6222e1dcb5e511acae1d78b8"), "brand_id" : 1, "brand_name" : "Honda" }
>
```

findOne()

In order to get one document that satisfies the search criteria, we can simply use the findOne() method on our chosen collection. If multiple documents satisfy the query, this method returns the first document according to the natural order which reflects the order of documents on the disk. If no documents satisfy the search criteria, the function returns null. The function takes the following form of syntax.

Syntax- db.{collection}.findOne({query}, {projection})

Example : db.sales.order_items.findOne({"quantity": 2})

```
> db.sales.order_items.findOne({"quantity": 2})
{
    "_id" : ObjectId("6222e0bab5e511acae1d78a8"),
    "order_id" : "ORD001",
    "product_id" : 1,
    "quantity" : 2,
    "list_price" : 50000
}>
```

Query Documents

a. **Specify Equality Condition**

db.production.products.find() - to show all the documents

db.production.products.find({"list_price": 30000}) - to show only documents which have “list_price” as 30000

```
> db.sales.customers.find()
{ "_id" : ObjectId("6222e05db5e511acae1d78a3"), "customer_id" : "Cus001", "first_name" : "Jay", "last_name" : "Mehta", "phone" : 1234567890, "email" : "jay@gmail.com", "street" : "T.P Road", "city" : "Mumbai", "state" : "Maharashtra", "zip_code" : 400072 }
{ "_id" : ObjectId("6222e05db5e511acae1d78a4"), "customer_id" : "Cus002", "first_name" : "Ruhi", "last_name" : "Singh", "phone" : 7894561230, "email" : "ruhi@yahoo.com", "street" : "M.G Chauk", "city" : "Mumbai", "state" : "Maharashtra", "zip_code" : 400072 }
{ "_id" : ObjectId("6222e05db5e511acae1d78a5"), "customer_id" : "Cus003", "first_name" : "Aria", "last_name" : "Josh", "phone" : 1245789630, "email" : "aria.josh@gmail.com", "street" : "JVM", "city" : "Gandhi Nagar", "state" : "Gujrat", "zip_code" : 401235 }
{ "_id" : ObjectId("6222e05db5e511acae1d78a6"), "customer_id" : "Cus004", "first_name" : "Mahi", "last_name" : "Kaur", "phone" : 4567890123, "email" : "kaur.mahi@hotmail.com", "street" : "J.V.L.R", "city" : "Mumbai", "state" : "Maharashtra", "zip_code" : 400072 }
{ "_id" : ObjectId("6222e05db5e511acae1d78a7"), "customer_id" : "Cus005", "first_name" : "Aditya", "last_name" : "Yadav", "phone" : 9638527410, "email" : "aditya@gmail.com", "street" : "Koliwada", "city" : "Pune", "state" : "Maharashtra", "zip_code" : 300075 }
> db.production.products.find({list_price: 30000})
{ "_id" : ObjectId("6222df37b5e511acae1d7899"), "product_id" : 2, "product_name" : "6KU Bikes", "brand_id" : 2, "category_id" : 4, "model_year" : 2000, "list_price" : 30000 }
{ "_id" : ObjectId("6222df37b5e511acae1d789a"), "product_id" : 3, "product_name" : "Bianchi", "brand_id" : 3, "category_id" : 2, "model_year" : 2002, "list_price" : 30000 }
>
```

b. Specify Conditions Using Query Operators

db.sales.customers.find() - to show all the documents

db.sales.customers.find({city: { \$in: ["Mumbai", "Pune"] } }) - to show all the documents where "city" is either "Mumbai" or "Pune"

```
> db.sales.customers.find()
{ "_id" : ObjectId("6222e05db5e511acae1d78a3"), "customer_id" : "Cus001", "first_name" : "Jay", "last_name" : "Mehta", "phone" : 1234567890, "email" : "jay@gmail.com", "street" : "T.P Road", "city" : "Mumbai", "state" : "Maharashtra", "zip_code" : 400072 }
{ "_id" : ObjectId("6222e05db5e511acae1d78a4"), "customer_id" : "Cus002", "first_name" : "Ruhi", "last_name" : "Singh", "phone" : 7894561230, "email" : "ruhi@yahoo.com", "street" : "M.G Chauk", "city" : "Mumbai", "state" : "Maharashtra", "zip_code" : 400072 }
{ "_id" : ObjectId("6222e05db5e511acae1d78a5"), "customer_id" : "Cus003", "first_name" : "Aria", "last_name" : "Josh", "phone" : 1245789630, "email" : "aria.josh@gmail.com", "street" : "JVM", "city" : "Gandhi Nagar", "state" : "Gujrat", "zip_code" : 401235 }
{ "_id" : ObjectId("6222e05db5e511acae1d78a6"), "customer_id" : "Cus004", "first_name" : "Mahi", "last_name" : "Kaur", "phone" : 4567890123, "email" : "kaur.mahi@hotmail.com", "street" : "J.V.L.R", "city" : "Mumbai", "state" : "Maharashtra", "zip_code" : 400072 }
{ "_id" : ObjectId("6222e05db5e511acae1d78a7"), "customer_id" : "Cus005", "first_name" : "Aditya", "last_name" : "Yadav", "phone" : 9638527410, "email" : "aditya@gmail.com", "street" : "Koliwada", "city" : "Pune", "state" : "Maharashtra", "zip_code" : 300075 }
> db.sales.customers.find({city: { $in: [ "Mumbai", "Pune" ] } })
{ "_id" : ObjectId("6222e05db5e511acae1d78a3"), "customer_id" : "Cus001", "first_name" : "Jay", "last_name" : "Mehta", "phone" : 1234567890, "email" : "jay@gmail.com", "street" : "T.P Road", "city" : "Mumbai", "state" : "Maharashtra", "zip_code" : 400072 }
{ "_id" : ObjectId("6222e05db5e511acae1d78a4"), "customer_id" : "Cus002", "first_name" : "Ruhi", "last_name" : "Singh", "phone" : 7894561230, "email" : "ruhi@yahoo.com", "street" : "M.G Chauk", "city" : "Mumbai", "state" : "Maharashtra", "zip_code" : 400072 }
{ "_id" : ObjectId("6222e05db5e511acae1d78a6"), "customer_id" : "Cus004", "first_name" : "Mahi", "last_name" : "Kaur", "phone" : 4567890123, "email" : "kaur.mahi@hotmail.com", "street" : "J.V.L.R", "city" : "Mumbai", "state" : "Maharashtra", "zip_code" : 400072 }
{ "_id" : ObjectId("6222e05db5e511acae1d78a7"), "customer_id" : "Cus005", "first_name" : "Aditya", "last_name" : "Yadav", "phone" : 9638527410, "email" : "aditya@gmail.com", "street" : "Koliwada", "city" : "Pune", "state" : "Maharashtra", "zip_code" : 300075 }
```

c. Specify AND Conditions

db.sales.order_items.find() - to show all the documents

db.sales.order_items.find({quantity: 2, list_price : { \$gt: 70000}}) - Here we are trying to find how sales order item documents for which quantity is 2 and list price is greater than 70000

```
> db.sales.order_items.find()
{ "_id" : ObjectId("6222e0bab5e511acae1d78a8"), "order_id" : "ORD001", "product_id"
: 1, "quantity" : 2, "list_price" : 5000 }
{ "_id" : ObjectId("6222e0bab5e511acae1d78a9"), "order_id" : "ORD002", "product_id"
: 2, "quantity" : 3, "list_price" : 9000 }
{ "_id" : ObjectId("6222e0bab5e511acae1d78aa"), "order_id" : "ORD003", "product_id"
: 3, "quantity" : 1, "list_price" : 3000 }
{ "_id" : ObjectId("6222e0bab5e511acae1d78ab"), "order_id" : "ORD004", "product_id"
: 4, "quantity" : 8, "list_price" : 36000 }
{ "_id" : ObjectId("6222e0bab5e511acae1d78ac"), "order_id" : "ORD005", "product_id"
: 5, "quantity" : 2, "list_price" : 10000 }
> db.sales.order_items.find({quantity: 2, list_price : { $gt: 70000}})
{ "_id" : ObjectId("6222e0bab5e511acae1d78ac"), "order_id" : "ORD005", "product_id"
: 5, "quantity" : 2, "list_price" : 10000 }
```

d. Specify OR Conditions

db.production.products.find() - to show all the documents

db.production.products.find({ \$or: [{ product_name: "Honda Superfast" }, { model_year : { \$lt: 2003 } }] }) - to show all the document which is either "product name" as "Honda Superfast" or "model year" is less than year 2003

Command Prompt - MONGO

```
> db.production.products.find()
{ "_id" : ObjectId("6222df37b5e511acae1d7898"), "product_id" : 1, "product_name" :
"Honda Superfast", "brand_id" : 1, "category_id" : 1, "model_year" : 1994, "list_pr
ice" : 25000 }
{ "_id" : ObjectId("6222df37b5e511acae1d7899"), "product_id" : 2, "product_name" :
"6KU Bikes", "brand_id" : 2, "category_id" : 4, "model_year" : 2000, "list_price" :
30000 }
{ "_id" : ObjectId("6222df37b5e511acae1d789a"), "product_id" : 3, "product_name" :
"Bianchi", "brand_id" : 3, "category_id" : 2, "model_year" : 2002, "list_price" : 3
0000 }
{ "_id" : ObjectId("6222df37b5e511acae1d789b"), "product_id" : 4, "product_name" :
"BMC Hybrid Bike", "brand_id" : 4, "category_id" : 3, "model_year" : 2009, "list_pr
ice" : 45000 }
{ "_id" : ObjectId("6222df37b5e511acae1d789c"), "product_id" : 5, "product_name" :
"Huffy Women Bike", "brand_id" : 5, "category_id" : 7, "model_year" : 2019, "list_p
rice" : 50000 }
> db.production.products.find({ $or: [ { product_name: "Honda Superfast" }, { model
_year : { $lt:
... 2003 } } ] })
{ "_id" : ObjectId("6222df37b5e511acae1d7898"), "product_id" : 1, "product_name" :
"Honda Superfast", "brand_id" : 1, "category_id" : 1, "model_year" : 1994, "list_pr
ice" : 25000 }
{ "_id" : ObjectId("6222df37b5e511acae1d7899"), "product_id" : 2, "product_name" :
"6KU Bikes", "brand_id" : 2, "category_id" : 4, "model_year" : 2000, "list_price" :
30000 }
{ "_id" : ObjectId("6222df37b5e511acae1d789a"), "product_id" : 3, "product_name" :
"Bianchi", "brand_id" : 3, "category_id" : 2, "model_year" : 2002, "list_price" : 3
0000 }
> -
```

3) Update Operations

Like create operations, update operations operate on a single collection, and they are atomic at a single document level. An update operation takes filters and criteria to select the documents you want to update.

You should be careful when updating documents, as updates are permanent and can't be rolled back. This applies to delete operations as well.

For MongoDB CRUD, there are three different methods of updating documents:

- db.collection.updateOne()
- db.collection.updateMany()
- db.collection.replaceOne()

1. updateOne()

We can update a currently existing record and change a single document with an update operation. To do this, we use the updateOne() method on a chosen collection. To update a document, we provide the method with two arguments: an update filter and an update action.

The update filter defines which items we want to update, and the update action defines how to update those items.

We first pass in the update filter. Then, we use the "\$set" key and provide the fields we want to update as a value.

This method will update the first record that matches the provided filter.

Example –**db.sales.staffs.find()** - to show current document in the system**db.sales.staffs.updateOne({first_name: "Pushpa"}, {\$set:{phone: 999999988}})** - update mobile number from 9999999999 to 999999988 for document name where name is " first_name " in collection

Command Prompt - MONGO

```
> db.production.products.find()
{ "_id" : ObjectId("6222df37b5e511acae1d7898"), "product_id" : 1, "product_name" :
"Honda Superfast", "brand_id" : 1, "category_id" : 1, "model_year" : 1994, "list_pr
ice" : 25000 }
{ "_id" : ObjectId("6222df37b5e511acae1d7899"), "product_id" : 2, "product_name" :
"6KU Bikes", "brand_id" : 2, "category_id" : 4, "model_year" : 2000, "list_price" :
30000 }
{ "_id" : ObjectId("6222df37b5e511acae1d789a"), "product_id" : 3, "product_name" :
"Bianchi", "brand_id" : 3, "category_id" : 2, "model_year" : 2002, "list_price" : 3
0000 }
{ "_id" : ObjectId("6222df37b5e511acae1d789b"), "product_id" : 4, "product_name" :
"BMC Hybrid Bike", "brand_id" : 4, "category_id" : 3, "model_year" : 2009, "list_pr
ice" : 45000 }
{ "_id" : ObjectId("6222df37b5e511acae1d789c"), "product_id" : 5, "product_name" :
"Huffy Women Bike", "brand_id" : 5, "category_id" : 7, "model_year" : 2019, "list_p
rice" : 50000 }
> db.production.products.find({ $or: [ { product_name: "Honda Superfast" }, { model
_year : { $lt:
... 2003 } } ] })
{ "_id" : ObjectId("6222df37b5e511acae1d7898"), "product_id" : 1, "product_name" :
"Honda Superfast", "brand_id" : 1, "category_id" : 1, "model_year" : 1994, "list_pr
ice" : 25000 }
{ "_id" : ObjectId("6222df37b5e511acae1d7899"), "product_id" : 2, "product_name" :
"6KU Bikes", "brand_id" : 2, "category_id" : 4, "model_year" : 2000, "list_price" :
30000 }
{ "_id" : ObjectId("6222df37b5e511acae1d789a"), "product_id" : 3, "product_name" :
"Bianchi", "brand_id" : 3, "category_id" : 2, "model_year" : 2002, "list_price" : 3
0000 }
> db.sales.staffs.find()
{ "_id" : ObjectId("6222e161b5e511acae1d78b2"), "staff_id" : 1, "first_name" : "Pus
hpaa", "last_name" : "Yadav", "email" : "pushpa@gmail.com", "phone" : 9999999999, "a
ctive" : "Yes", "store_id" : 1, "manager_id" : 1 }
{ "_id" : ObjectId("6222e161b5e511acae1d78b3"), "staff_id" : 2, "first_name" : "Sad
iksha", "last_name" : "Singh", "email" : "sadiksha@gmail.com", "phone" : 8888888888
, "active" : "Yes", "store_id" : 2, "manager_id" : 1 }
{ "_id" : ObjectId("6222e161b5e511acae1d78b4"), "staff_id" : 3, "first_name" : "Pri
ya", "last_name" : "Nadar", "email" : "priya@gmail.com", "phone" : 7777777777, "act
ive" : "Yes", "store_id" : 3, "manager_id" : 1 }
> db.sales.staffs.updateOne({first_name: "Pushpa"}, {$set:{phone: 999999988}})
{ "acknowledged" : true, "matchedCount" : 1, "modifiedCount" : 1 }
>
```

2. updateMany()

updateMany() allows us to update multiple items by passing in a list of items, just as we did when inserting multiple items. This update operation uses the same syntax for updating a single document.

Example –**db.sales.orders.find()** - to show current document in the system**db.sales.orders.updateMany({shipped_date:44377}, {\$set: {shipped_date: "1-July-2021"}})** – with the help of this command we are updating "shipped_date" to "1-July-2021" where shipped_date is 44377

Command Prompt - MONGO

```
{ "acknowledged" : true, "matchedCount" : 1, "modifiedCount" : 1 }
> db.sales.orders.find()
{ "_id" : ObjectId("6222e0f7b5e511acae1d78ad"), "order_id" : "ORD001", "customer_id
" : "Cus001", "order_status" : "Completed", "order_date" : 43992, "shipped_date" :
43994, "store_id" : 1, "staff_id" : 1 }
{ "_id" : ObjectId("6222e0f7b5e511acae1d78ae"), "order_id" : "ORD002", "customer_id
" : "Cus002", "order_status" : "Completed", "order_date" : 44221, "shipped_date" :
44227, "store_id" : 2, "staff_id" : 2 }
{ "_id" : ObjectId("6222e0f7b5e511acae1d78af"), "order_id" : "ORD003", "customer_id
" : "Cus003", "order_status" : "Completed", "order_date" : 44306, "shipped_date" :
44314, "store_id" : 2, "staff_id" : 2 }
{ "_id" : ObjectId("6222e0f7b5e511acae1d78b0"), "order_id" : "ORD004", "customer_id
" : "Cus004", "order_status" : "Pending", "order_date" : 44367, "shipped_date" : 44
377, "store_id" : 3, "staff_id" : 3 }
{ "_id" : ObjectId("6222e0f7b5e511acae1d78b1"), "order_id" : "ORD005", "customer_id
" : "Cus005", "order_status" : "Pending", "order_date" : 44367, "shipped_date" : 44
377, "store_id" : 1, "staff_id" : 1 }
> db.sales.orders.updateMany({shipped_date:44377}, {$set: {shipped_date: "1-July-20
21"}})
{ "acknowledged" : true, "matchedCount" : 2, "modifiedCount" : 2 }
>
```

3. replaceOne()

The replaceOne() method is used to replace a single document in the specified collection. replaceOne() replaces the entire document, meaning fields in the old document not contained in the new will be lost.

4) Delete Operations

Delete operations operate on a single collection, like update and create operations. Delete operations are also atomic for a single document. You can provide delete operations with filters and criteria in order to specify which documents you would like to delete from a collection. The filter options rely on the same syntax that read operations utilize.

MongoDB has two different methods of deleting records from a collection:

- db.collection.deleteOne()
- db.collection.deleteMany()

1. deleteOne()

deleteOne() is used to remove a document from a specified collection on the MongoDB server. A filter criteria is used to specify the item to delete. It deletes the first record that matches the provided filter.

Example –

db.production.brands.find() - to show current documents in collection

db.production.brands.deleteOne({brand_id:6}) - delete the document where "brand_id" is 6

```
> db.production.brands.find()
{ "_id" : ObjectId("6222e1dcb5e511acae1d78b8"), "brand_id" : 1, "brand_name" : "Hon
da" }
{ "_id" : ObjectId("6222e1dcb5e511acae1d78b9"), "brand_id" : 2, "brand_name" : "6KU
Bikes" }
{ "_id" : ObjectId("6222e1dcb5e511acae1d78ba"), "brand_id" : 3, "brand_name" : "Bia
nchi" }
{ "_id" : ObjectId("6222e1dcb5e511acae1d78bb"), "brand_id" : 4, "brand_name" : "BMC
" }
{ "_id" : ObjectId("6222e1dcb5e511acae1d78bc"), "brand_id" : 5, "brand_name" : "Huf
fy" }
> db.production.brands.deleteOne({brand_id:6})
{ "acknowledged" : true, "deletedCount" : 0 }
```

2. deleteMany()

deleteMany() is a method used to delete multiple documents from a desired collection with a single delete operation. A list is passed into the method and the individual items are defined with filter criteria as in deleteOne().

Example –

db.production.brands.find() - to show current documents in collection

db.production.brands.deleteMany({brand_id: {\$gt: 5}}) - delete documents for which brand id is greater than 5

```
Command Prompt - MONGO
> db.production.brands.deleteOne({brand_id:6})
{ "acknowledged" : true, "deletedCount" : 0 }
> db.production.brands.find()
{ "_id" : ObjectId("6222e1dcb5e511acae1d78b8"), "brand_id" : 1, "brand_name" : "Hon
da" }
{ "_id" : ObjectId("6222e1dcb5e511acae1d78b9"), "brand_id" : 2, "brand_name" : "6KU
Bikes" }
{ "_id" : ObjectId("6222e1dcb5e511acae1d78ba"), "brand_id" : 3, "brand_name" : "Bia
nchi" }
{ "_id" : ObjectId("6222e1dcb5e511acae1d78bb"), "brand_id" : 4, "brand_name" : "BMC
" }
{ "_id" : ObjectId("6222e1dcb5e511acae1d78bc"), "brand_id" : 5, "brand_name" : "Huf
fy" }
> db.production.brands.deleteMany({brand_id: {$gt: 5}})
{ "acknowledged" : true, "deletedCount" : 0 }
```

Aim: Joins, Sorting, Subqueries using HiveQL

JOINS

JOIN is a clause that is used for combining specific fields from two tables by using values common to each one. It is used to combine records from two or more tables in the database.

There are different types of joins given as follows:

- JOIN
- LEFT OUTER JOIN
- RIGHT OUTER JOIN
- FULL OUTER JOIN
- **JOIN**

JOIN clause is used to combine and retrieve the records from multiple tables. JOIN is same as OUTER JOIN in SQL. A JOIN condition is to be raised using the primary keys and foreign keys of the tables.

➤ **LEFT OUTER JOIN**

The HiveQL LEFT OUTER JOIN returns all the rows from the left table, even if there are no matches in the right table. This means, if the ON clause matches 0 (zero) records in the right table, the JOIN still returns a row in the result, but with NULL in each column from the right table. A LEFT JOIN returns all the values from the left table, plus the matched values from the right table, or NULL in case of no matching JOIN predicate.

➤ **RIGHT OUTER JOIN**

The HiveQL RIGHT OUTER JOIN returns all the rows from the right table, even if there are no matches in the left table. If the ON clause matches 0 (zero) records in the left table, the JOIN still returns a row in the result, but with NULL in each column from the left table.

A RIGHT JOIN returns all the values from the right table, plus the matched values from the left table, or NULL in case of no matching join predicate

➤ **FULL OUTER JOIN**

The HiveQL FULL OUTER JOIN combines the records of both the left and the right outer tables that fulfil the JOIN condition. The joined table contains either all the records from both the tables, or fills in NULL values for missing matches on either side.

SUB QUERIES:

A Query present within a Query is known as a sub query. The main query will depend on the values returned by the subqueries.

Subqueries can be classified into two types

- Subqueries in **FROM** clause
- Subqueries in **WHERE** clause

When to use:

- To get a particular value combined from two column values from different tables
- Dependency of one table values on other tables
- Comparative checking of one column values from other tables

SORTING

The SORT BY syntax is similar to the syntax of ORDER BY in SQL language.

Hive supports SORT BY which sorts the data per reducer. The difference between "order by" and "sort by" is that the former guarantees total order in the output while the latter only guarantees ordering of the rows within a reducer. If there are more than one reducer, "sort by" may give partially ordered final results.

Hive uses the columns in SORT BY to sort the rows before feeding the rows to a reducer. The sort order will be dependent on the column types. If the column is of numeric type, then the sort order is also in numeric order. If the column is of string type, then the sort order will be lexicographical order

Before the perform the practical first perform 3 steps of the given following

1. sudo /home/cloudera/cloudera-manager --force --express

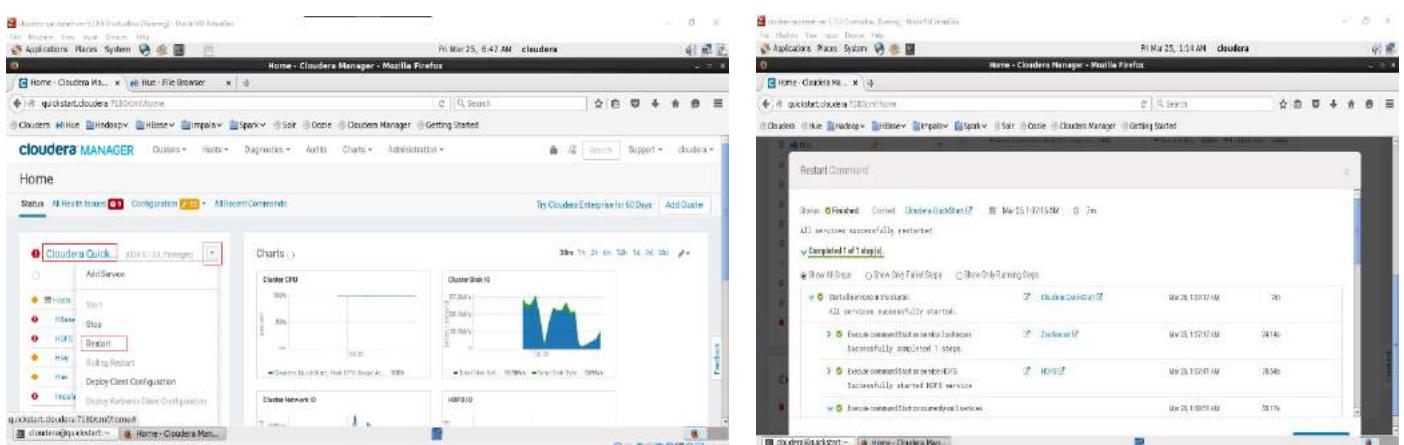
```
[QuickStart] Waiting for Cloudera Manager API...
[QuickStart] Starting Cloudera Manager agent...
[QuickStart] Configuring deployment...
Submitted jobs: 41
[QuickStart] Deploying client configuration...
Submitted jobs: 42
[QuickStart] Starting Cloudera Management Service...
Submitted jobs: 58
[QuickStart] Enabling Cloudera Manager daemons on boot...

Success! You can now log into Cloudera Manager from the QuickStart VM's browser:
http://quickstart.cloudera:7180
Username: cloudera
Password: cloudera
```

[cloudera@quickstart ~]\$ sudo -u hdfs hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

22/03/25 01:04:22 WARN ipc.Client: Failed to connect to server: quickstart.cloudera/10.0.2.15:8828: try once and fail.
java.net.ConnectException: Connection refused
at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
at sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:239)
at org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout.java:206)
at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:538)
at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:494)
at org.apache.hadoop.ipc.Client\$Connection.setupConnection(Client.java:648)
at org.apache.hadoop.ipc.Client\$Connection.access\$3000(Client.java:744)
at org.apache.hadoop.ipc.Client\$Connection.access\$3000(Client.java:399)

2. Start all services



3. sudo -u hdfs hadoop dfsadmin -safemode leave

```
cloudera@quickstart-vm-5130-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Fri Mar 25, 1:16 AM cloudera@quickstart:~
```

```
sudo -u hdfs hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

safe mode is OFF
[cloudera@quickstart ~]$
```

```
cloudera@quickstart-vm-5130-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Fri Mar 25, 1:16 AM cloudera@quickstart:~
```

```
sudo -u hdfs hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

safe mode is OFF
[cloudera@quickstart ~]$
```

```
cloudera@quickstart-vm-5130-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Fri Mar 25, 1:16 AM cloudera@quickstart:~
```

```
sudo -u hdfs hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

safe mode is OFF
[cloudera@quickstart ~]$
```

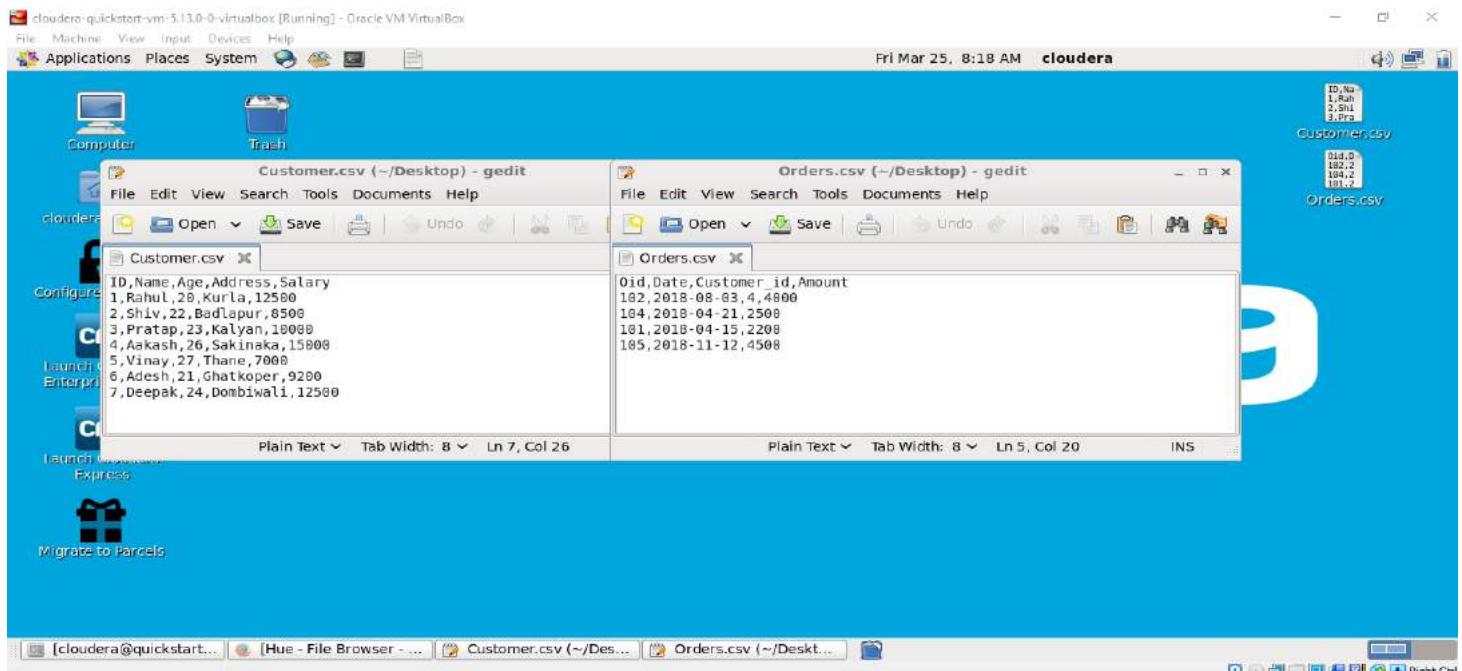
First we will create the **Customer.csv** and **Order.csv** file.

Customer.csv

ID	Name	Age	Address	Salary
1	Rahul	20	Kurla	12500
2	Shiv	22	Badlapur	8500
3	Pratap	23	Kalyan	10000
4	Aakash	26	Sakinaka	15000
5	Vinay	27	Thane	7000
6	Adesh	21	Ghatkoper	9200
7	Deepak	24	Dombiwali	12500

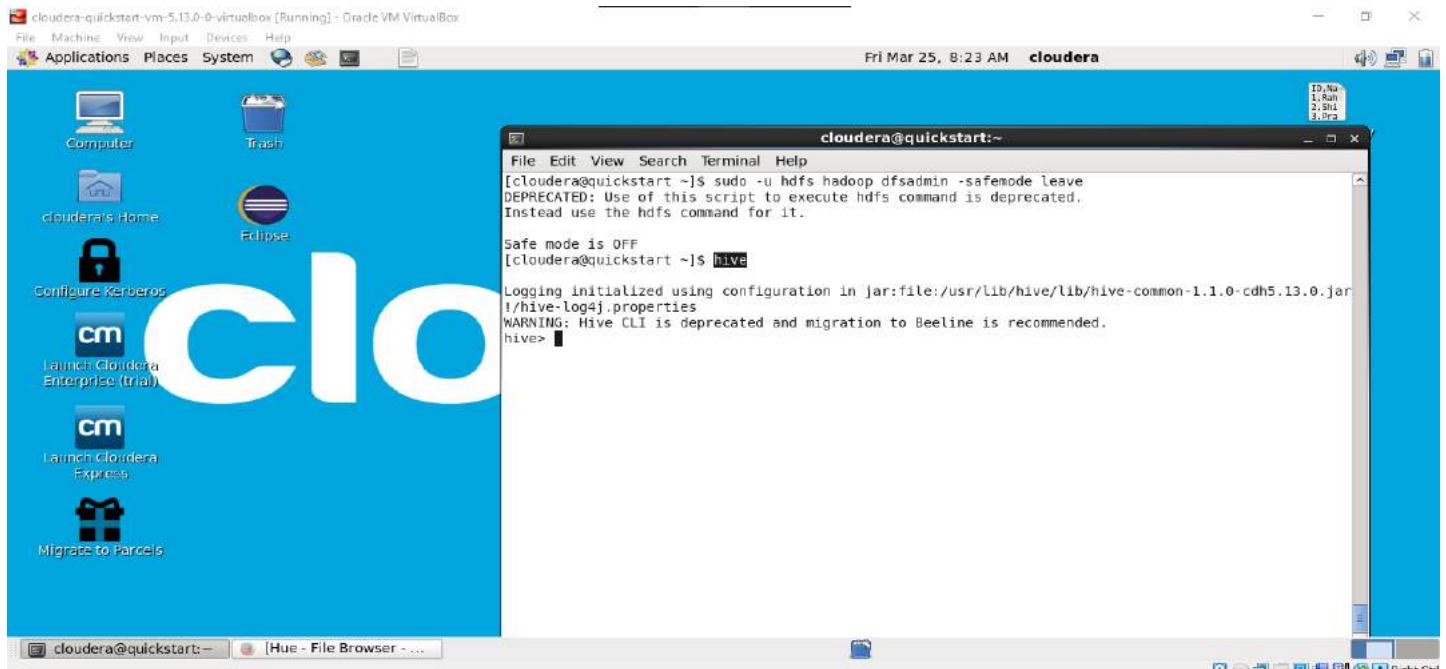
Orders.csv

Oid	Date	Customer_id	Amount
102	2018-08-03	4	4000
104	2018-04-21	1	2500
101	2018-04-15	1	2200
105	2018-11-12	2	4500



Open the terminal, now we use **hive** command to enter the **hive shell prompt** and in hive shell we could execute all of the hive commands.

1. **hive**



cloudera@quickstart:~

```

at org.antlr.runtime.DFA.predict(DFA.java:116)
at org.apache.hadoop.hive.parser.HiveParser.ddlStatement(HiveParser.java:2503)
at org.apache.hadoop.hive ql.parse.HiveParser.execStatement(HiveParser.java:1589)
at org.apache.hadoop.hive ql.parse.HiveParser.statement(HiveParser.java:1065)
at org.apache.hadoop.hive ql.parse.ParseDriver.parse(ParseDriver.java:281)
at org.apache.hadoop.hive ql.parse.ParseDriver.parse(ParseDriver.java:166)
at org.apache.hadoop.hive ql.Driver.compile(Driver.java:522)
at org.apache.hadoop.hive ql.Driver.compileInternal(Driver.java:1356)
at org.apache.hadoop.hive ql.Driver.runInternal(Driver.java:1473)
at org.apache.hadoop.hive ql.Driver.run(Driver.java:1285)
at org.apache.hadoop.hive ql.Driver.run(Driver.java:1275)
at org.apache.hadoop.hive cli.CliDriver.processLocalCmd(CliDriver.java:226)
at org.apache.hadoop.hive cli.CliDriver.processCmd(CliDriver.java:175)
at org.apache.hadoop.hive cli.CliDriver.processLine(CliDriver.java:389)
at org.apache.hadoop.hive cli.CliDriver.executeDriver(CliDriver.java:781)
at org.apache.hadoop.hive cli.CliDriver.run(CliDriver.java:699)
at org.apache.hadoop.hive cli.CliDriver.main(CliDriver.java:634)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:606)
at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 1:5 cannot recognize input near 'show' 'database' '<EOF>' in ddl statement
hive> show databases;
OK
default
Time taken: 3.497 seconds, Fetched: 1 row(s)
hive> 
```

Now we will be creating a new database named as **rjc_joins** using below command,

2. **create database rjc_joins;**

cloudera@quickstart:~

```

at org.apache.hadoop.hive ql.parse.HiveParser.statement(HiveParser.java:1065)
at org.apache.hadoop.hive ql.parse.ParseDriver.parse(ParseDriver.java:281)
at org.apache.hadoop.hive ql.parse.ParseDriver.parse(ParseDriver.java:166)
at org.apache.hadoop.hive ql.Driver.compile(Driver.java:522)
at org.apache.hadoop.hive ql.Driver.compileInternal(Driver.java:1356)
at org.apache.hadoop.hive ql.Driver.runInternal(Driver.java:1473)
at org.apache.hadoop.hive ql.Driver.run(Driver.java:1285)
at org.apache.hadoop.hive ql.Driver.run(Driver.java:1275)
at org.apache.hadoop.hive cli.CliDriver.processLocalCmd(CliDriver.java:226)
at org.apache.hadoop.hive cli.CliDriver.processCmd(CliDriver.java:175)
at org.apache.hadoop.hive cli.CliDriver.processLine(CliDriver.java:389)
at org.apache.hadoop.hive cli.CliDriver.executeDriver(CliDriver.java:781)
at org.apache.hadoop.hive cli.CliDriver.run(CliDriver.java:699)
at org.apache.hadoop.hive cli.CliDriver.main(CliDriver.java:634)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:606)
at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 1:5 cannot recognize input near 'show' 'database' '<EOF>' in ddl statement
T
hive> show databases;
OK
default
Time taken: 3.497 seconds, Fetched: 1 row(s)
hive> create database rjc_joins;
OK
Time taken: 0.371 seconds
hive> 
```

And then showing the databases.

show databases;

cloudera@quickstart:~

```

File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~
```

cloudera@quickstart:~

```

at org.apache.hadoop.hive ql.Driver.runInternal(Driver.java:1473)
at org.apache.hadoop.hive ql.Driver.run(Driver.java:1285)
at org.apache.hadoop.hive ql.Driver.run(Driver.java:1275)
at org.apache.hadoop.hive cli.CliDriver.processLocalCmd(CliDriver.java:226)
at org.apache.hadoop.hive cli.CliDriver.processCmd(CliDriver.java:175)
at org.apache.hadoop.hive cli.CliDriver.processLine(CliDriver.java:389)
at org.apache.hadoop.hive cli.CliDriver.executeDriver(CliDriver.java:781)
at org.apache.hadoop.hive cli.CliDriver.run(CliDriver.java:699)
at org.apache.hadoop.hive cli.CliDriver.main(CliDriver.java:634)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:606)
at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
FAILED: ParseException line 1:5 cannot recognize input near 'show' 'database' '<EOF>' in ddl statement
T
hive> show databases;
OK
default
Time taken: 3.497 seconds, Fetched: 1 row(s)
hive> create database rjc_joins;
OK
Time taken: 0.371 seconds
hive> [show databases];
OK
default
rjc_joins
Time taken: 0.061 seconds, Fetched: 2 row(s)
hive> 
```

Now to work inside this database we use below command;

3. use rjc_joins;

```

cloudera@quickstart:~$ hive> use rjc_joins;
OK
Time taken: 0.249 seconds
hive>

```

Now we will create two tables in one table we will load the **Customer.csv** file and in the other table we will load **Orders.csv** file.

4. create table customers(ID int, Name string, Age int, Address string, Salary float)

row format delimited

fields terminated by ','

tblproperties("skip.header.line.count"="1");

```

cloudera@quickstart:~$ hive> create table customers(ID int, Name string, Age int, Address string, Salary float)
> row format delimited
> fields terminated by ','
> tblproperties("skip.header.line.count"="1");
OK
Time taken: 2.289 seconds
hive>

```

show tables;

```

hive> show tables;
OK
customers
Time taken: 0.759 seconds, Fetched: 1 row(s)
hive>

```

Now we will see the schema of the table using describe command, **describe customers;**

describe customers;

```
hive> describe customers;
OK
id          int
name        string
age         int
address    string
salary      float
Time taken: 0.782 seconds, Fetched: 5 row(s)
hive> [REDACTED]
```

Now loading data in the **customers** table from **Customer.csv** file which present inside **/home/cloudera/<dir_name>** directory.

5. load data local inpath "/home/cloudera/Desktop/Customer.csv" into table customers;

```
hive> load data local inpath "/home/cloudera/Desktop/Customer.csv" into table customers
> ;
Loading data to table rjc_joins.customers
Table rjc_joins.customers stats: (numFiles=1, totalSize=203)
OK
Time taken: 4.24 seconds
hive> [REDACTED]
```

select * from customers;

```
hive> select * from customers;
OK
1    Rahul  20    Kurla   12500.0
2    Shiv   22    Badlapur 8500.0
3    Pratap 23    Kalyan  10000.0
4    Aakash 26    Sakinaka 15000.0
5    Vinay  27    Thane   7000.0
6    Adesh  21    Ghatkoper 9200.0
7    Deepak 24    Dombiwali 12500.0
NULL NULL  NULL  NULL  NULL
Time taken: 1.7 seconds, Fetched: 8 row(s)
hive> [REDACTED]
```

Creating a second table named as **orders** using below command,

6. create table orders(oid int, odate date, cid int, amount float)

row format delimited

fields terminated by ','

tblproperties("skip.header.line.count"="1");

```
hive> create table orders(oid int, odate date, cid int, amount float)
> row format delimited
> fields terminated by ','
> tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.558 seconds
hive> [REDACTED]
```

Now we will see the schema of the table using **describe** command,

describe orders;

```
hive> describe orders
> ;
OK
oid          int
odate       date
cid          int
amount      float
Time taken: 0.443 seconds, Fetched: 4 row(s)
hive> [REDACTED]
```

Now loading data in the **orders** table from **Orders.csv** file which present inside `/home/cloudera/<dir_name>` directory.

7. load data local inpath "/home/cloudera/Desktop/Orders.csv" into table orders;

```
hive> load data local inpath "/home/cloudera/Desktop/Orders.csv" into table orders
>;
Loading data to table rjc_joins.orders
Table rjc_joins.orders stats: [numFiles=1, totalSize=111]
OK
Time taken: 1.422 seconds
hive>
```

`select * from orders;`

```
hive> select * from orders;
OK
102 2018-08-03 4 4000.0
104 2018-04-21 2500 NULL
101 2018-04-15 2200 NULL
105 2018-11-12 4500 NULL
Time taken: 0.381 seconds, Fetched: 4 row(s)
hive> truncate table orders;
OK
Time taken: 1.605 seconds
hive> load data local inpath "/home/cloudera/Desktop/Orders.csv" into table orders;
Loading data to table rjc_joins.orders
Table rjc_joins.orders stats: [numFiles=1, numRows=0, totalSize=116, rawDataSize=0]
OK
Time taken: 3.16 seconds
hive> select * from orders;
OK
102 2018-08-03 4 4000.0
104 2018-04-21 1 2500.0
101 2018-04-15 1 2200.0
105 2018-11-12 2 4500.0
Time taken: 0.851 seconds, Fetched: 4 row(s)
hive>
```

JOIN

Now first we apply the normal joins on the two tables using below command, we want to retrieve customer id, name, age from customers table and amount from the orders table and join perform on id of the customers and orders table.

8. select c.id, c.name, c.age, o.amount

from customers c JOIN orders o
on (c.id = o.cid);

```
hive> select c.id, c.name, c.age, o.amount
> from customers c JOIN orders o
> on (c.id = o.cid);
Query ID = cloudera_20220325085858_6ac28064-2d67-4188-9d19-e62b7ee0c754
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20220325085858_6ac28064-2d67-4188-9d19-e62b7ee0c754.log
```

Browser - ...

Mapreduce task is performed

```
cloudera-quickstart-vm-5.13.0-0-virtualBox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Terminal Help
cloudera@quickstart:~$ File Edit View Search Terminal Help
37654
2022-03-25 08:59:32 Dump the side-table for tag: 1 with group count: 3 into file: file:/tmp/cloud
era/2a5ce55c-e04e-490f-a2e1-11dacd82dec6/hive_2022-03-25_08-58-49_024_4092038695331254827-1/_local-10
003/HashTable-Stage-3/MapJoin-mapfile01--.hashtable
2022-03-25 08:59:33 Uploaded 1 File to: file:/tmp/cloudera/2a5ce55c-e04e-490f-a2e1-11dacd82dec6/hive_2022-03-25_08-58-40_024_4092038695331254827-1/_local-10003/HashTable-Stage-3/MapJoin-mapfile01--.hashtable (338 bytes)
2022-03-25 08:59:33 End of local task; Time Taken: 13.051 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1648216557393_0005, Tracking URL = http://quickstart.cloudera:8888/proxy/applicati
on_1648216557393_0005
Kill Command = user/lib/hadoop/bin/hadoop job -kill job_1648216557393_0005
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-03-25 09:00:27,665 Stage-3 map = 0% reduce = 0%
2022-03-25 09:01:12,537 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 6.25 sec
MapReduce Total cumulative CPU time: 6 seconds 250 msec
Ended Job = job_1648216557393_0005
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1 Cumulative CPU: 6.25 sec HDFS Read: 7098 HDFS Write: 72 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 250 msec
OK
1 Rahul 20 2500.0
2 Rahul 20 2200.0
3 Shiv 22 4500.0
4 Akash 26 4000.0
Time taken: 155.387 seconds, Fetched: 4 row(s)
hive>
```

LEFT OUTER JOIN

The HiveQL LEFT OUTER JOIN returns all the rows from the left table, even if there are no matches in the right table. This means, if the ON clause matches 0 (zero) records in the right table, the JOIN still returns a row in the result, but with NULL in each column from the right table.

A LEFT JOIN returns all the values from the left table, plus the matched values from the right table, or NULL in case of no matching JOIN predicate.

**9. select c.id, c.name, o.amount,o.odate
from customers c LEFT OUTER JOIN orders o
on (c.id = o.cid);**

```
hive> select c.id, c.name, o.amount,o.odate
> from customers c LEFT OUTER JOIN orders o
> on (c.id = o.cid);
Query ID = cloudera_20220325090202_75101a32-531d-4e2c-9d17-1e716cf37995
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20220325090202_75101a32-531d-4e2c-9d17-1e716cf37995.log
2022-03-25 09:03:25      Starting to launch local task to process map join;      maximum memory = 1195
37664
```

Mapreduce task is performed

```
File Edit View Search Terminal Help
2022-03-25 09:07:43 Uploaded 1 File to: file:/tmp/cloudera/2a5ce55c-e04e-490f-a2e1-11dacd82dec6/hive_2022-03-25_09-07-11_649_1068820263093827760-1-local-10803/HashTable-Stage-3/MapJoin-mapfile21--.hashtable (350 bytes)
2022-03-25 09:07:43 End of local task; Time Taken: 8.349 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job: job_1648216557393_0007, Tracking URL: http://quickstart.cloudera:8088/proxy/application_1648216557393_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648216557393_0007
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-03-25 09:08:29.715 Stage-3 map = 0%, reduce = 0%
2022-03-25 09:09:01.145 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 5.11 sec
MapReduce Total cumulative CPU time: 5 seconds 110 msec
Ended Job = job_1648216557393_0007
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1 Cumulative CPU: 5.11 sec HDFS Read: 7156 HDFS Write: 162 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 110 msec
OK
1 Rahul 2500.0 2018-04-21
1 Rahul 2200.0 2018-04-15
2 Shiv 4500.0 2018-11-12
3 Pratap NULL NULL
4 Aakash 4000.0 2018-08-03
5 Vinay NULL NULL
6 Adesh NULL NULL
7 Deepak NULL NULL
Time taken: 111.942 seconds, Fetched: 8 rows(s)
hive>
```

RIGHT OUTER JOIN

The HiveQL RIGHT OUTER JOIN returns all the rows from the right table, even if there are no matches in the left table. If the ON clause matches 0 (zero) records in the left table, the JOIN still returns a row in the result, but with NULL in each column from the left table.

A RIGHT JOIN returns all the values from the right table, plus the matched values from the left table, or NULL in case of no matching join predicate.

**10. select c.id, c.name, o.amount,o.odate
from customers c RIGHT OUTER JOIN orders o
on (c.id = o.cid);**

```
hive> select c.id, c.name, o.amount,o.odate
> from customers c RIGHT OUTER JOIN orders o
> on (c.id = o.cid);
Query ID = cloudera_20220325090909_f1d30aa1-55a3-4835-ab4c-d7263d18f42d
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20220325090909_f1d30aa1-55a3-4835-ab4c-d7263d18f42d.log
2022-03-25 09:10:20      Starting to launch local task to process map join;      maximum memory = 1195
37664
2022-03-25 09:10:26      Dump the side-table for tag: 0 with group count: 7 into file: file:/tmp/cloud
era/2a5ce55c-e04e-490f-a2e1-11dacd82dec6/hive_2022-03-25_09-09-58_105_406453300447223904-1-local-10
003/HashTable-Stage-3/MapJoin-mapfile30--.hashtable
```

Mapreduce task is performed

```

cloudera@quickstart:~$ hadoop jar /tmp/cloudera/2a5ce55c-e04e-490f-a2e1-11dacd82dec6/hive_2022-03-25_09-09-58_105_4064533008447223904-1-local-10003/HashTable-Stage-3/MapJoin-mapfile30--.hashtable
2022-03-25 09:10:27 Uploaded 1 File to: file:/tmp/cloudera/2a5ce55c-e04e-490f-a2e1-11dacd82dec6/hive_2022-03-25_09-09-58_105_4064533008447223904-1-local-10003/HashTable-Stage-3/MapJoin-mapfile30--.hashtable (437 bytes)
2022-03-25 09:10:27 End of local task; Time Taken: 6.154 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job
Number of Reduce Tasks is set to 0 since there's no reduce operator
Starting Job = job_1648216557393_0008, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648216557393_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648216557393_0008
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 0
2022-03-25 09:11:17,568 Stage-3 map = 0%, reduce = 0%
2022-03-25 09:12:00,647 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 5.4 sec
MapReduce Total cumulative CPU time: 5 seconds 400 msec
Ended Job = job_1648216557393_0008
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1 Cumulative CPU: 5.4 sec HDFS Read: 7160 HDFS Write: 104 SUCCESS
Total MapReduce CPU Time Spent: 5 seconds 400 msec
OK
4 Aakash 4000.0 2018-08-03
1 Rahul 2500.0 2018-04-21
1 Rahul 2200.0 2018-04-15
2 Shiv 4500.0 2018-11-12
Time taken: 125.03 seconds, Fetched: 4 row(s)
hive>

```

Now we will be using the concept of **subqueries** for finding the second largest salary from the customers table.

Sub Queries

A Query present within a Query is known as a sub query. The main query will depend on the values returned by the subqueries.

Subqueries can be classified into two types

- Subqueries in FROM clause
- Subqueries in WHERE clause

11. select max(salary) from customers where customers.salary not in(select max(salary) from customers);

```

hive> select max(salary) from customers where customers.salary not in(select max(salary) from customers);
Warning: Map Join MAPJOIN[62][bigTable=customers] in task 'Stage-8:MAPRED' is a cross product
Warning: Shuffle Join JOIN[24][tables = [customers, sq_1_notin_nullcheck]] in Stage 'Stage-1:MAPRED' is a cross product
Query ID = cloudera_20220325091414_55e79d66-e08d-419a-a4f1-7eaf6b4960fc
Total jobs = 7
Launching Job 1 out of 7
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1648216557393_0009, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648216557393_0009/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648216557393_0009

```

Mapreduce task is performed

```

cloudera@quickstart:~$ hadoop jar /tmp/cloudera/2a5ce55c-e04e-490f-a2e1-11dacd82dec6/hive_2022-03-25_09-09-58_105_4064533008447223904-1-local-10003/HashTable-Stage-3/MapJoin-mapfile30--.hashtable
2022-03-25 09:10:27 Dump the side-table for tag: 0 with group count: 7 into file: file:/tmp/cloudera/2a5ce55c-e04e-490f-a2e1-11dacd82dec6/hive_2022-03-25_09-09-58_105_4064533008447223904-1-local-10003/HashTable-Stage-3/MapJoin-mapfile30--.hashtable
2022-03-25 09:10:27 Uploaded 1 File to: file:/tmp/cloudera/2a5ce55c-e04e-490f-a2e1-11dacd82dec6/hive_2022-03-25_09-09-58_105_4064533008447223904-1-local-10003/HashTable-Stage-3/MapJoin-mapfile30--.hashtable (437 bytes)
2022-03-25 09:10:27 End of local task; Time Taken: 6.154 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job
Number of Reduce Tasks is set to 0 since there's no reduce operator
Starting Job = job_1648216557393_0013, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648216557393_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648216557393_0013
Hadoop job information for Stage-3: number of mappers: 1; number of reducers: 1
2022-03-25 09:22:06,415 Stage-3 map = 0%, reduce = 0%
2022-03-25 09:22:46,757 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 3.23 sec
2022-03-25 09:23:14,132 Stage-3 map = 100%, reduce = 100%, Cumulative CPU 8.68 sec
MapReduce Total cumulative CPU time: 8 seconds 680 msec
Ended Job = job_1648216557393_0013
MapReduce Jobs Launched:
Stage-Stage-3: Map: 1 Reduce: 1 Cumulative CPU: 9.32 sec HDFS Read: 7332 HDFS Write: 117 SUCCESS
Stage-Stage-3: Map: 1 Reduce: 1 Cumulative CPU: 10.58 sec HDFS Read: 8700 HDFS Write: 114 SUCCESS
Stage-Stage-8: Map: 1 Cumulative CPU: 5.61 sec HDFS Read: 5349 HDFS Write: 243 SUCCESS
Stage-Stage-6: Map: 1 Cumulative CPU: 5.86 sec HDFS Read: 3116 HDFS Write: 117 SUCCESS
Stage-Stage-3: Map: 1 Reduce: 1 Cumulative CPU: 8.88 sec HDFS Read: 4715 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 40 seconds 630 msec
OK
12500.0
Time taken: 536.151 seconds, Fetched: 1 row(s)
hive>

```

As we can see from the above output the second largest salary is **12500.00**.

Sorting

The SORT BY syntax is similar to the syntax of ORDER BY in SQL language.

Hive supports **SORT BY** which sorts the data per reducer. The difference between "order by" and "sort by" is that the former guarantees total order in the output while the latter only guarantees ordering of the rows within a reducer. If there are more than one reducer, "sort by" may give partially ordered final results.

Hive uses the columns in SORT BY to sort the rows before feeding the rows to a reducer. The sort order will be dependent on the column types. If the column is of numeric type, then the sort order is also in numeric order. If the column is of string type, then the sort order will be lexicographical order.

LIMIT can be used to minimize sort time.

Now finding the **fourth largest salary** from the customers table using **Sort by** clause.

12. select salary from customers sort by salary desc limit 4;

It will give the only 4 records in the output after sorting them in descending order. This is not a complete syntax only we are showing what output it will give.

```

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Customer.csv (~/Desktop) - Edit
File Edit View Search Tools Documents Help
Customer.csv
File Edit View Terminal Help
2022-03-25 09:12:06,415 Stage-3 map = 0%, reduce = 0%
2022-03-25 09:22:46,757 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 3.23 sec
2022-03-25 09:23:14,132 Stage-3 map = 100%, reduce = 100%, Cumulative CPU 8.68 sec
MapReduce Total cumulative CPU time: 8 seconds 680 msec
Ended Job = job_1648216557393_0013
MapReduce Jobs Launched:
Stage-Stage-4: Map: 1 Reduce: 1 Cumulative CPU: 9.92 sec HDFS Read: 7333 HDFS Write: 117 SUCCESS
Stage-Stage-5: Map: 1 Reduce: 1 Cumulative CPU: 10.56 sec HDFS Read: 8700 HDFS Write: 114 SUCCESS
Stage-Stage-6: Map: 1 Cumulative CPU: 5.61 sec HDFS Read: 5349 HDFS Write: 243 SUCCESS
Stage-Stage-7: Map: 1 Cumulative CPU: 5.88 sec HDFS Read: 5116 HDFS Write: 117 SUCCESS
Stage-Stage-8: Map: 1 Reduce: 1 Cumulative CPU: 8.68 sec HDFS Read: 4715 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 40 seconds 630 msec
OK
12508.0
Time taken: 536.151 seconds, Fetched: 1 row(s)
hive> select salary from customers sort by salary desc limit 4;
Query ID = cloudera_20220325092626_a53cd5a-bcf3-4f4b-9011-c96534a186c6
Total jobs = 2
Launching Job 1 out of 2
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1648216557393_0014, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648216557393_0014/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648216557393_0014
[...]

```

Mapreduce task is performed

```

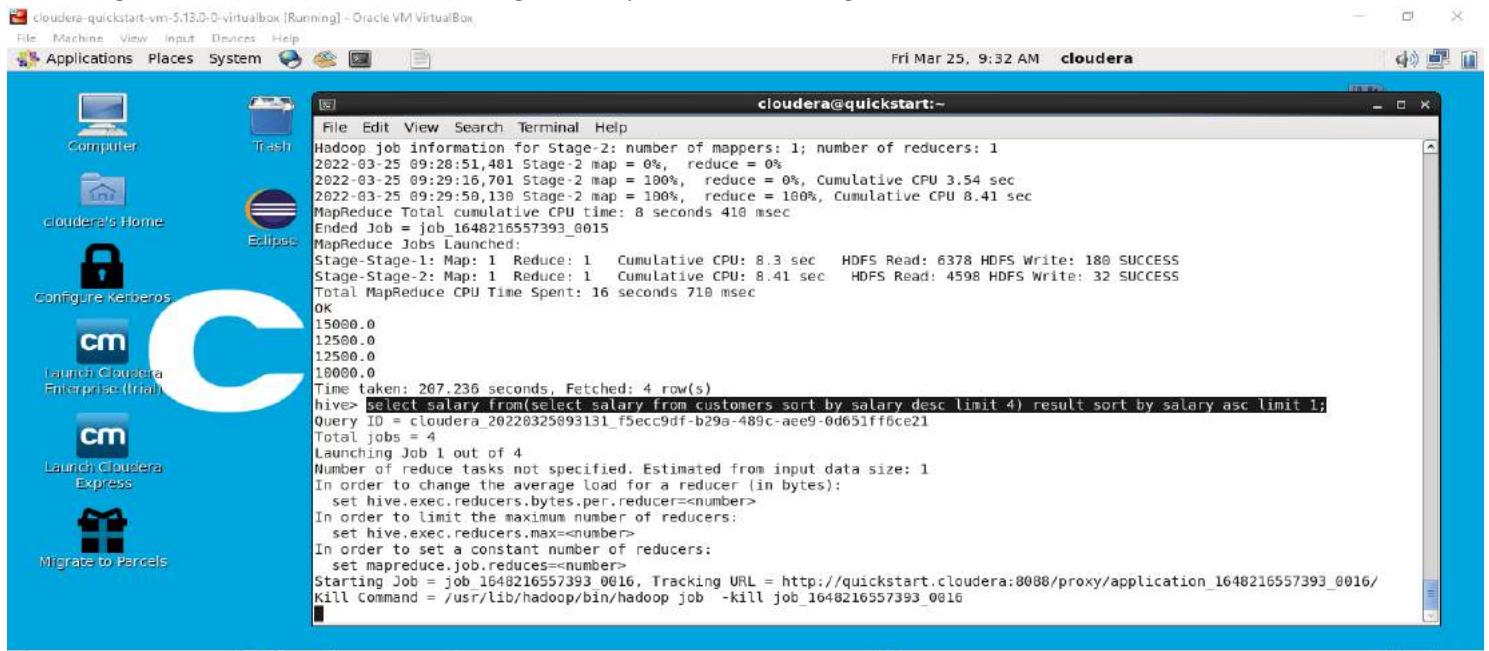
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Customer.csv (~/Desktop) - Edit
File Edit View Terminal Help
MapReduce Total cumulative CPU time: 8 seconds 300 msec
Ended Job = job_1648216557393_0014
Launching Job 0 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1648216557393_0015, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648216557393_0015/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648216557393_0015
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-03-25 09:28:51,481 Stage-2 map = 0%, reduce = 0%
2022-03-25 09:29:16,701 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 3.54 sec
2022-03-25 09:29:50,139 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 8.41 sec
MapReduce Total cumulative CPU time: 8 seconds 410 msec
Ended Job = job_1648216557393_0015
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.3 sec HDFS Read: 6378 HDFS Write: 180 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 8.41 sec HDFS Read: 4598 HDFS Write: 32 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 710 msec
OK
13000.0
12500.0
12000.0
11500.0
Time taken: 287.236 seconds, Fetched: 4 row(s)
hive> [...]

```

Now what records which we have got by executing the above queries now we will use this query as subqueries and we will now sort them in ascending order to find fourth largest salary of customer table.

13. select salary from(select salary from customers sort by salary desc limit 4) result sort by salary asc limit 1;

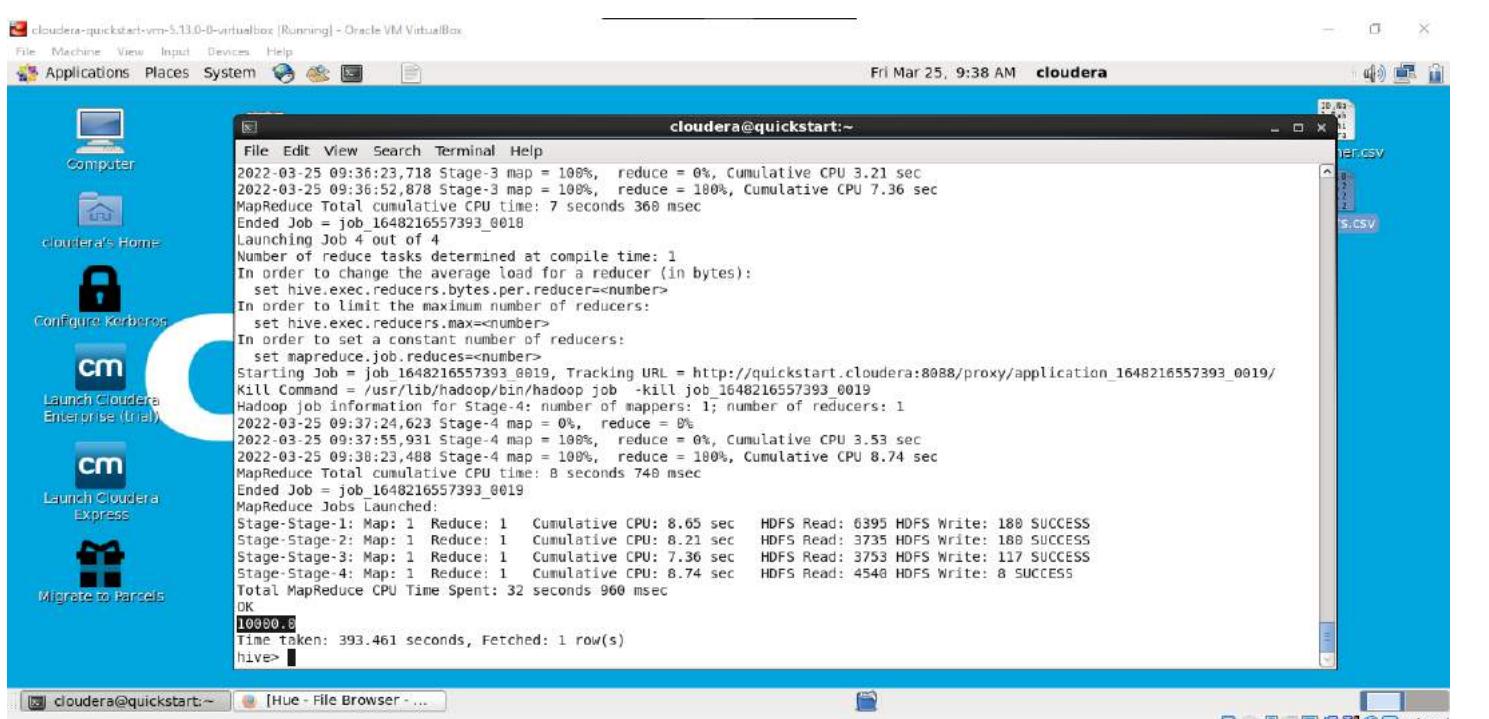
Now whatever result we get from subquery we will store them in result table and then it will sort the result table in ascending order and as we want fourth largest salary so we are limiting it to 1.



```

File Edit View Search Terminal Help
Hadoop job information for Stage-2: number of mappers: 1; number of reducers: 1
2022-03-25 09:28:51,481 Stage-2 map = 0%, reduce = 0%
2022-03-25 09:29:15,701 Stage-2 map = 100%, reduce = 0%, Cumulative CPU 3.54 sec
2022-03-25 09:29:50,130 Stage-2 map = 100%, reduce = 100%, Cumulative CPU 8.41 sec
MapReduce Total cumulative CPU time: 8 seconds 410 msec
Ended Job = job_1648216557393_0015
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.3 sec HDFS Read: 6378 HDFS Write: 180 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 8.41 sec HDFS Read: 4598 HDFS Write: 32 SUCCESS
Total MapReduce CPU Time Spent: 16 seconds 710 msec
OK
15000.0
12500.0
12500.0
10000.0
Time taken: 207.236 seconds, Fetched: 4 row(s)
hive> select salary from(select salary from customers sort by salary desc limit 4) result sort by salary asc limit 1;
Query ID = cloudera_20220325093131_f5ecc9df-b29a-489c-aee9-0d651ff6ce21
Total jobs = 4
Launching Job 1 out of 4
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<numbers>
Starting Job = job_1648216557393_0016, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648216557393_0016/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648216557393_0016

```



```

File Edit View Search Terminal Help
2022-03-25 09:38:23,718 Stage-3 map = 100%, reduce = 0%, Cumulative CPU 3.21 sec
2022-03-25 09:38:52,878 Stage-3 map = 100%, reduce = 100%, Cumulative CPU 7.36 sec
MapReduce Total cumulative CPU time: 7 seconds 360 msec
Ended Job = job_1648216557393_0018
Launching Job 4 out of 4
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1648216557393_0019, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648216557393_0019/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648216557393_0019
Hadoop job information for Stage-4: number of mappers: 1; number of reducers: 1
2022-03-25 09:37:24,623 Stage-4 map = 0%, reduce = 0%
2022-03-25 09:37:55,931 Stage-4 map = 100%, reduce = 0%, Cumulative CPU 3.53 sec
2022-03-25 09:38:23,480 Stage-4 map = 100%, reduce = 100%, Cumulative CPU 8.74 sec
MapReduce Total cumulative CPU time: 8 seconds 740 msec
Ended Job = job_1648216557393_0019
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.65 sec HDFS Read: 6395 HDFS Write: 180 SUCCESS
Stage-Stage-2: Map: 1 Reduce: 1 Cumulative CPU: 8.21 sec HDFS Read: 3735 HDFS Write: 180 SUCCESS
Stage-Stage-3: Map: 1 Reduce: 1 Cumulative CPU: 7.36 sec HDFS Read: 3753 HDFS Write: 117 SUCCESS
Stage-Stage-4: Map: 1 Reduce: 1 Cumulative CPU: 8.74 sec HDFS Read: 4540 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 32 seconds 960 msec
OK
10000.0
Time taken: 393.461 seconds, Fetched: 1 row(s)
hive> 
```

Now we got the fourth largest salary i.e. 10000.0 as an output.

Aim: Querying, Sorting, Aggregating data using HiveQL

What is HIVE:

Hive is a data warehouse system which is used to analyze structured data. It is built on the top of Hadoop. It was developed by Facebook.

Hive provides the functionality of reading, writing, and managing large datasets residing in distributed storage. It runs SQL like queries called HQL (Hive query language) which gets internally converted to MapReduce jobs.

Using Hive, we can skip the requirement of the traditional approach of writing complex MapReduce programs. Hive supports Data Definition Language (DDL), Data Manipulation Language (DML), and User Defined Functions (UDF).

Features of Hive:

These are the following features of Hive:

- ❖ Hive is fast and scalable.
 - ❖ It provides SQL-like queries (i.e., HQL) that are implicitly transformed to MapReduce or Spark jobs.
 - ❖ It is capable of analyzing large datasets stored in HDFS.
 - ❖ It allows different storage types such as plain text, RCFile, and HBase.
 - ❖ It uses indexing to accelerate queries.
 - ❖ It can operate on compressed data stored in the Hadoop ecosystem.
 - ❖ It supports user-defined functions (UDFs) where user can provide its functionality.

Limitations of Hive

- Hive is not capable of handling real-time data.
 - It is not designed for online transaction processing.
 - Hive queries contain high latency.

Before the perform the practical first perform 3 steps of the given following

1. **sudo /home/cloudera/cloudera-manager --force --express**

```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System cloudera@quickstart:~ Fri Mar 25, 1:15 AM cloudera
[QuickStart] Waiting for Cloudera Manager API...
[QuickStart] Starting Cloudera Manager agent...
[QuickStart] Configuring deployment...
Submitted jobs: 41
[QuickStart] Deploying client configuration...
Submitted jobs: 42
[QuickStart] Starting Cloudera Management Service...
Submitted jobs: 50
[QuickStart] Enabling Cloudera Manager daemons on boot...

Success! You can now log into Cloudera Manager from the QuickStart VM's browser:
http://quickstart.cloudera:7180

Username: cloudera
Password: cloudera

[cloudera@quickstart ~]$ sudo -u hdfs hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

22/03/25 01:04:22 WARN ipc.Client: Failed to connect to server: quickstart.cloudera:10.0.2.15:8820; try once and fail.
java.net.ConnectException: Connection refused
    at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
    at sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:739)
    at org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout.java:206)
    at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:538)
    at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:494)
    at org.apache.hadoop.ipc.Client$Connection.setupConnection(Client.java:648)
    at org.apache.hadoop.ipc.Client$Connection.setupIOstreams(Client.java:744)
    at org.apache.hadoop.ipc.Client$Connection.access$3000(Client.java:390)

cloudera@quickstart:~ | Home - Cloudera Man...
```

2. Start all services

The left screenshot shows the Cloudera Manager dashboard with several service status indicators. The right screenshot shows a detailed log of restart commands, with a summary at the top stating "All services successfully restarted".

3. sudo -u hdfs hadoop dfsadmin -safemode leave

```

cloudera@quickstart:~$ sudo -u hdfs hadoop dfsadmin -safemode leave
at org.apache.hadoop.ipc.ClientsConnection.setupConnection(Client.java:648)
at org.apache.hadoop.ipc.ClientsConnection.setupIOstreams(Client.java:744)
at org.apache.hadoop.ipc.ClientsConnection.access$3000(Client.java:386)
at org.apache.hadoop.ipc.ClientsConnection.getConnection(Client.java:1557)
at org.apache.hadoop.ipc.Client.call(Client.java:1488)
at org.apache.hadoop.ipc.Client.call(Client.java:1441)
at org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngine.java:230)
at com.sun.proxy.$Proxy16.setSafeMode(Unknown Source)
at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.setSafeMode(ClientNamenodeProtocolTranslatorPB.java:681)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:606)
at org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryInvocationHandler.java:288)
at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:104)
at com.sun.proxy.$Proxy17.setSafeMode(Unknown Source)
at org.apache.hadoop.hdfs.DFSClient.setSafeMode(DFSClient.java:2648)
at org.apache.hadoop.hdfs.DistributedFileSystem.setSafeMode(DistributedFileSystem.java:1182)
at org.apache.hadoop.hdfs.DistributedFileSystem.setSafeMode(DistributedFileSystem.java:1166)
at org.apache.hadoop.hdfs.tools.DFSAdmin.setSafeMode(DFSAdmin.java:576)
at org.apache.hadoop.hdfs.tools.DFSAdmin.run(DFSAdmin.java:1856)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:84)
at org.apache.hadoop.hdfs.tools.DFSAdmin.main(DFSAdmin.java:2032)

safeMode: Call From quickstart.cloudera:18.8.2.15 to quickstart.cloudera:8020 failed on connection exception: java.net.ConnectException: Connection refused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
[cloudera@quickstart ~]$ sudo -u hdfs hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Safe mode is OFF
[cloudera@quickstart ~]$

```

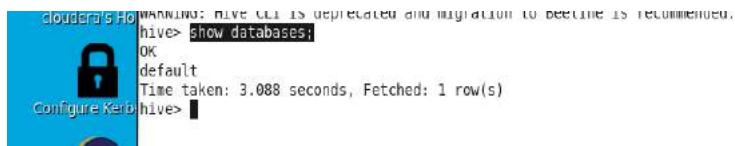
1. Open the terminal, Now we use hive command to enter the hive shell prompt and in hive shell we could execute all of the hive commands.

hive



2. Now we will see the databases which are already existing using below command.

show databases;



Note: If remove database already present in hive then **drop database <database_name> cascade;**

3. Creating a database name 'RJC' using below command.

create database RJC;

```
Configure Kerberos
hive> create database RJC;
OK
Time taken: 0.178 seconds
hive> [REDACTED]
```



4. So if we want to see whether this RJC database is created or not we will use below command,

show databases;

```
Launch Cloud Enterprise (trial)
hive> show databases;
OK
default
rjc
Time taken: 0.036 seconds, Fetched: 2 row(s)
hive> [REDACTED]
```



5. Now we want to check whether we have any tables inside this rjc database or not. So first we will move to this database rjc using below command,

use rjc;

```
cm
Launch Cloud Express
hive> use rjc;
OK
Time taken: 0.054 seconds
hive> [REDACTED]
```



6. Now we have moved inside this rjc database. Now we will check out which are the tables available using below command,

show tables;

```
cm
Launch Cloud Express
hive> show tables;
OK
Time taken: 0.106 seconds
hive> [REDACTED]
```



So as we can see from the above output it give us OK as output because there are no tables created inside this rjc database. So as there are no tables we did not get anything in the output we simply got as OK.

No will drop this 'rjc' database using below command we do not mention here cascade because there are no tables in this and so there are no data available or present inside this rjc database.

drop database rjc;

and to check whether the data base dropped or not using below command,

show databases;

```

hive> drop database rjc;
OK
Time taken: 0.264 seconds
hive> show databases;
OK
default
Time taken: 0.024 seconds, Fetched: 1 row(s)
hive> [REDACTED]

```

[cloudera@quickstart:~] [Home - Cloudera Man...]

7. Creating a database "rjc".

create database rjc;

Now let's move to this database using below command so now all the work we will do or perform it should be done within this rjc database.

use rjc;

```

cloudera's Home
hive> create database RJC;
OK
Time taken: 0.086 seconds
hive> show databases;
OK
default
rjc
Time taken: 0.023 seconds, Fetched: 2 row(s)
hive> use rjc;

```

8. Now we will create a table inside this rjc database named as employee using below command,

create table employee(ID int, name string, salary float, age int)

Note: After this we will not put **semicolon**, When we will be loading the data from some existing csv file or maybe some other text files so we have to mention that how that data has to be loaded here. We are simply creating the schema of the table with some certain fields or attributes along with their datatypes and then I'm mentioning

row format delimited

fields terminated by ',';

```

hive> create table employee(ID int, name string, salary float, age int)
   > row format delimited
   > fields terminated by ',';
OK
Time taken: 0.425 seconds
hive> show tables;
OK
employee
Time taken: 0.043 seconds, Fetched: 1 row(s)
hive> [REDACTED]

```

[cloudera@quickstart:~] [Home - Cloudera Man...]

row format delimited means , every record is present in one row and **fields terminated by ','** and **fields are terminated by comma**. So as soon as it encounters one comma so that means that one is the value of some field and after comma it is encountering abc so that abc is the value of some another field. By default it is a "tab" character that means fields are separated by "tab".

9. So now we will see the schema of the table using below command,

describe employee;

```

hive> describe employee;
OK
id          int
name        string
salary      float
age         int
Time taken: 0.167 seconds, Fetched: 4 row(s)
hive> [REDACTED]

```

[cloudera@quickstart:~] [Home - Cloudera Man...]

It will give different fields of table employee along with their respective datatypes.

10. By default the internal table would store in the warehouse directory of hive. Whereas the external tables are available in the hdfs. And if we drop the internal table so then the table data and the metadata associated with that table will be deleted from the hdfs. Whereas when we drop the external table then only the metadata associated with that table will be deleted whereas the table data will be untouched by hive as it would be residing in the hdfs and it would be outside the warehouse directory of the hive. So Now we will check how the table which we will be created is internal table or external table using below command,

describe formatted employee;

By default hive creates Internal table or **Managed Table**.

11. Now we will create the external table using below command,

Note: external

Do this

create external table employee2(ID int, name string, salary float, age int)

row format delimited

fields terminated by ','

stored as textfile;

```
Time taken: 0.749 seconds
hive> create external table employee2(ID int, name string, salary float, age int)
      > row format delimited
      > fields terminated by(',')
      > stored as textfile;
OK
Time taken: 0.19 seconds
hive> 
```

Don't this

create table employee2(ID int, name string, salary float, age int)

row format delimited

fields terminated by ','

stored as textfile;

```
hive> create table employee2(ID int, name string, salary float, age int)
      > row format delimited
      > fields terminated by(',')
      > stored as textfile;
OK
Time taken: 0.158 seconds
hive> 
```

12. Checking the schema of the table using below command,

describe employee2;

```
hive> describe employee2;
OK
id          int
name        string
salary      float
age         int
Time taken: 0.147 seconds, Fetched: 4 row(s)
hive> 
```

13. So Now we will check how the table which we will be created is internal table or external table using below command,

describe formatted employee2;

Don't

```
cloudera@quickstart:~$ describe formatted employee2;
+-----+-----+
| name | type |
+-----+-----+
| id   | int   |
| name | string |
| age  | int   |
+-----+-----+
2 rows selected, 0.001 seconds, Fetched: 30 rows
```

It is an Managed table. (If not write external)

Do

```
cloudera@quickstart:~$ describe formatted employee2;
+-----+-----+
| name | type |
+-----+-----+
| id   | int   |
| name | string |
| age  | int   |
+-----+-----+
2 rows selected, 0.001 seconds, Fetched: 30 rows

Detailed Table Information
Name: employee2
Owner: cloudera
CreateTime: Fri Mar 25 10:32:01 PDT 2011
LastAccessTime: Fri Mar 25 10:32:01 PDT 2011
ProtectMode: None
StorageType: HDFS
Location: https://quickstart.cloudera:8080/user/mive/warehouse/employee2
Table Type: EXTERNAL TABLE
Table Privileges:
Table Properties:
    comment: None
    transient: TRUE
    version: 201103250002
```

It is an External table.

14. So whatever we have done in terminal the same thing we can see in the browser as well.

Open the browser → click on Hue

The screenshot shows the Cloudera Manager dashboard. In the top navigation bar, 'Hue' is highlighted. On the left sidebar, under 'Cloudera QuickStart', the 'Hue' service is listed with a green icon and the status '1 healthy'. The main area displays three charts: 'Cluster CPU' (CPU usage over time), 'Cluster Disk IO' (disk I/O rates), and 'HDFS IO' (HDFS block replication progress). Below the charts, there is a section titled 'Cloudera QuickStart' showing 'Total Bytes Received Across'.

Then click on Query->Editor->Hive

The screenshot shows the Hue Editor interface. At the top, the URL is 'quickstart.cloudera:8888/hue/editor?type=hive'. The main area is titled 'Hive' and contains a text input field with placeholder text: 'example: SELECT * FROM tablename; or press CTRL + space'. Below the input field are tabs for 'Query History' and 'Saved Queries'. To the right, there is a sidebar titled 'Assistant' with a 'Hive' dropdown menu containing various SQL functions like Aggregate, Analytic, Collection, etc. The bottom of the screen shows the terminal prompt '[cloudera@quickstart:~]' and the browser title 'Hue - Editor - Mozilla Firefox'.

Write the query in query editor. Here we are showing the databases by using below command,

show databases;

To execute -> CTRL + ENTER

It will give the list of databases which are present.

The screenshot shows the Hue Editor interface in Mozilla Firefox. The left sidebar under 'Hive' shows two databases: 'default' and 'rjc'. The main query editor window contains the command 'show databases;'. The results pane shows a table with one column 'database_name' containing two rows: '1 default' and '2 rjc'. To the right of the results is a sidebar with a 'Functions' section containing a list of Hive functions.

There are only two databases present i.e. default and rjc which we created earlier. We can also see the list of databases in left side corner. And after clicking on the database name here it is "rjc" it will give the list of all tables present inside "rjc" database.

Click on the rjc database

This screenshot is identical to the previous one, but the 'rjc' database is selected in the left sidebar. The query editor still shows 'show databases;', and the results table remains the same. The sidebar's function list is also visible.

The screenshot shows the Hue Editor interface in Mozilla Firefox. The left sidebar shows a tree view of tables under 'rjc'. The main area displays a Hive query:

```
show databases;
```

The results pane shows two databases: 'default' and 'rjc'. A sidebar on the right lists various Hive functions.

We can also Preview the sample data. Here it is showing blank because we have not inserted anything or data inside this employee table.

The screenshot shows the Hue Editor interface in Mozilla Firefox. The left sidebar shows a tree view of tables under 'rjc'. The main area displays a Hive query:

```
select * from rjc.employee;
```

The results pane shows a message: 'Done. 0 results.' Below the results, the query history shows previous queries like 'use rjc' and 'select * from employee'.

When we click on employee table it will give the fields of employee table.

The screenshot shows the Hue Editor interface in Mozilla Firefox. The left sidebar shows a tree view of tables under 'rjc'. A modal window is open for the 'employee' table, showing its columns:

Name	Type
1 id	Int
2 name	string
3 salary	float
4 age	Int

The main query editor area has a syntax error message: '1:0 cannot recognize input near'. The bottom status bar shows recent actions: '21 minutes ago employee2 formatted describe' and '76 minutes ago describe formatted employee2'.

cloudera@quickstart-sm-5:13:0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Home - Cloudera Manager Hue - Editor quickstart.cloudera:8888/hue/editor?editor=11
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
HUE
Tables (2) rjc employee employee2
Columns Details Sample
Columns (4)
Name Type
1 Id int
2 name string
3 salary float
4 age int
21 minutes ago employee2 formatted describe
24 minutes ago describe formatted employee2
cloudera@quickstart:~ Hue - Editor - Mozilla Firefox

Above if not created **external** table then first drop the table

drop table employee2;

```
Time taken: 0.749 seconds
hive> create external table employee2(ID int, name string, salary float, age int)
> row format delimited
> fields terminated by ','
> stored as textfile;
OK
Time taken: 0.19 seconds
hive> 
```

If view the properties **describe** command give the properties of employee2 table.

describe formatted employee2;

cloudera@quickstart-sm-5:13:0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Home - Cloudera Manager Hue - Editor quickstart.cloudera:8888/hue/editor?editor=15&type=hive
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
HUE
Tables (2) rjc employee employee2
Query col_name data_type
7 NULL
8 # Detailed Table Information NULL
9 Database: rjc
10 Owner: cloudera
11 CreateTime: Sat Mar 26 00:33:01 PDT 2022
12 LastAccessTime: UNKNOWN
13 Protect Mode: None
14 Retention: 0
15 Location: hdfs://quickstart.cloudera:8020/user/hive/warehouse/rjc.db/employee2
16 Table Type: EXTERNAL_TABLE
17 Table Parameters: NULL
18 EXTERNAL
cloudera@quickstart:~ Hue - Editor - Mozilla Firefox

```
cloudera@quickstart-sm-5:13:0-0-virtualbox [Running] - Oracle VM VirtualBox  
File Machine View Input Devices Help  
Applications Places System  
cloudera@quickstart:~  
File Edit View Search Terminal Help  
id int  
name string  
salary float  
age int  
# Detailed Table Information  
Database: rjc  
Owner: cloudera  
CreateTime: Sat Mar 26 00:33:01 PDT 2022  
LastAccessTime: UNKNOWN  
Protect Mode: None  
Retention: 0  
Location: hdfs://quickstart.cloudera:8020/user/hive/warehouse/rjc.db/employee2  
Table Type: EXTERNAL_TABLE  
Table Parameters:  
EXTERNAL TRUE  
transient_lastDdlTime 1648279981  
# Storage Information  
Serde Library: org.apache.hadoop.hive.serde2.lazy.LazySimpleSerde  
InputFormat: org.apache.hadoop.mapred.TextInputFormat  
OutputFormat: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat  
Compressed: No  
Num Buckets: 3  
Bucket Columns: []  
Sort Columns: []  
Storage Desc Params:  
field.delim  
serialization.format  
Time taken: 0.17 seconds. Fetched: 31 row(s)
hive> 
```

15. Creating a new external table named as employee3 in the specific location using below command,

create external table employee3(ID int, name string, salary float, age int)

row format delimited

fields terminated by ','

location '/user/cloudera/vj';

It will first create 'vj' directory inside the /user/cloudera and then inside 'vj' the employee3 table get stored.

```
> location '/user/cloudera/vj';
FAILED: Execution Error, return code 1 from org.apache.hadoop.hive.ql.exec.DDLTask. AlreadyExistsException(message:Table employee2 already exists)
hive> create external table employee3(ID int, name string, salary float, age int)
> row format delimited
> fields terminated by ','
> location '/user/cloudera/vj';
OK
Time taken: 0.174 seconds
hive>
```

16. To see the schema of the employee3 table we use below command,

describe employee3;

```
hive> describe employee3;
OK
id          int
name        string
salary      float
age         int
Time taken: 0.16 seconds, Fetched: 4 row(s)
hive>
```

17. Then switch to browser and Refresh and select the rjc database and refresh, now it will display the employee3 table along with other two tables which we have created earlier.

describe formatted employee3;

The screenshot shows a terminal window titled 'cloudera@quickstart:~'. The user has run the command 'create external table employee3(ID int, name string, salary float, age int) row format delimited fields terminated by ',' location '/user/cloudera/vj';'. The response indicates that the table 'employee2' already exists, so the command failed. The terminal also shows the time taken and the number of rows fetched.

Here we will see the properties of employee3 table here we can also see the location of the table where it is stored.

18. Now move to terminal and listing out all the tables using below command;

show tables;

```
hive> show tables;
OK
employee
employee2
employee3
Time taken: 0.025 seconds, Fetched: 3 row(s)
hive>
```

19. ALTER COMMANDS

If changing the name of the employee3 table to emptable using below command,

alter table employee3 RENAME TO emptable;

20. First we will see the fields of employee3 then we will add new column as surname in employee3 using below command,

alter table employee3 add columns(surname string);

```
hive> describe employee3;
OK
id          int
name        string
salary      float
age         int
Time taken: 0.15 seconds, Fetched: 4 row(s)
hive> alter table employee3 add columns(surname string);
OK
Time taken: 0.265 seconds
hive> describe employee3;
OK
id          int
name        string
salary      float
age         int
surname    string
Time taken: 0.186 seconds, Fetched: 5 row(s)
hive>
```

21. Now we will change field name of the emptable to first_name using alter command,

Syntax: **alter table <table_name> change <old_coln_name> <new_coln_name> string;**

alter table employee3 change name first_name string;

```
hive> alter table employee3 change name first name string;
OK
Time taken: 0.572 seconds
hive> describe employee3;
OK
id          int
first name    string
salary      float
age         int
surname    string
Time taken: 0.151 seconds, Fetched: 5 row(s)
hive>
```

Loading the data in the table

22. Before loading the data in the table we will first create the csv file. Now open the new terminal , using ls command list out all the directories --> change the directory to document directory --> use ls command to list all the files present inside the document folder or directory

The screenshot shows a desktop environment with a terminal window and a file manager window. The terminal window is titled 'cloudera@quickstart:~' and shows the command 'ls' being run, listing various files and directories including 'enterprise-deployment.json', 'kerberos', 'lib', and 'public'. The file manager window shows a file named 'Student.csv'.

```
cloudera@quickstart:~$ ls
enterprise-deployment.json  Music  Templates
Downloads  express-deployment.json  parcels  Videos
departments.java  dpt.java  kerberos  Pictures  workspace
Desktop  eclipse  lib  Public
[cloudera@quickstart ~]$
```

```
First_name    string
salary       float
age         int
surname    string
Time taken: 0.151 seconds, Fetched: 5 row(s)
hive>
```

23. Now creating new file as Student.csv using below command,

gedit Student.csv

As soon as we hit enter it will create a Student.csv file as it was not existing earlier.

1,Rahul,Hadoop,20

2,Shiv,Java,22

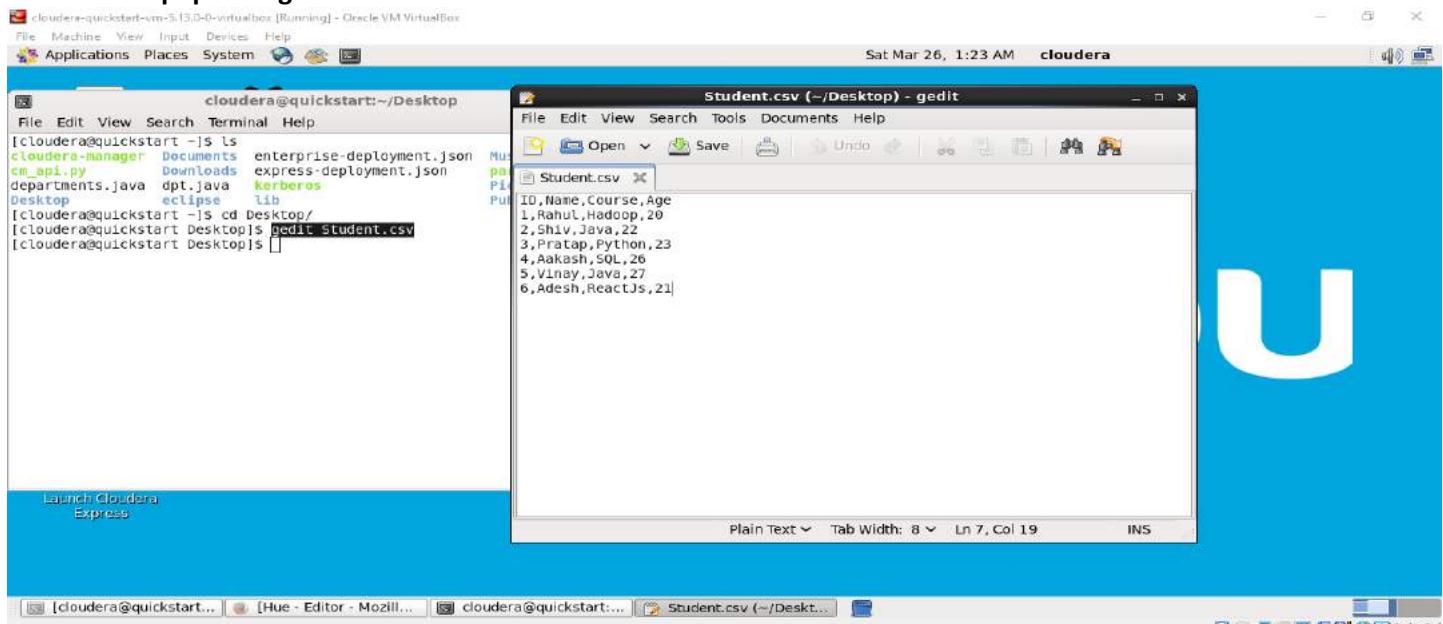
3,Pratap,Python,23

4,Aakash,SQL,26

5,Vinay,Java,27

6,Adesh,ReactJs,21

Now we are populating the Student.csv file with some data and save this file.



24. Creating a new database as rjcstudent.

create database rjcstudent;

```
hive> create database rjcstudent;
OK
Time taken: 0.109 seconds
hive> show databases;
OK
default
rjc
rjcstudent
Time taken: 0.022 seconds, Fetched: 3 row(s)
hive>
```

25. Using rjcstudent database.

use rjcstudent;

```
hive> show databases;
OK
default
rjc
rjcstudent
Time taken: 0.022 seconds, Fetched: 3 row(s)
hive> use rjcstudent;
OK
Time taken: 0.036 seconds
hive>
```

26. Creating new table student inside rjcstudent database.

```
create table student(ID int, Name string, Age int)
```

```
partitioned by(Course string)
```

```
row format delimited
```

```
fields terminated by ',';
```

```
hive> create table student(ID int, Name string, Age int)
> partitioned by(Course string)
> row format delimited
> fields terminated by ',';
OK
Time taken: 0.516 seconds
hive> ■
```

27. To see the structure or schema of the table,

```
describe student;
```

```
hive> describe student;
OK
id                  int
name                string
age                 int
course              string

# Partition Information
# col_name          data_type      comment
course              string
Time taken: 0.243 seconds, Fetched: 9 row(s)
```

28. Loading data in the student table from Student.csv file which we have created in document directory. Here we are partitioning based on course = 'Hadoop'.

```
load data local inpath "/home/cloudera/Desktop/Student.csv" into table student
```

```
partition(Course="Hadoop");
```

```
hive> load data local inpath "/home/cloudera/Desktop/Student.csv" into table student
> partition(Course="Hadoop");
Loading data to table rjcstudent.student partition (course=Hadoop)
Partition rjcstudent.student{course=Hadoop} stats: [numFiles=1, numRows=0, totalSize=122, rawDataSize=0]
OK
Time taken: 2.081 seconds
hive> ■
```

It is partitioning based on Hadoop.

```
select * from student;
```

```
hive> select * from student;
OK
NULL    Name     NULL    Hadoop
1       Rahul    NULL    Hadoop
2       Shiv     NULL    Hadoop
3       Pratap   NULL    Hadoop
4       Aakash   NULL    Hadoop
5       Vinay    NULL    Hadoop
6       Adesh    NULL    Hadoop
Time taken: 0.509 seconds, Fetched: 7 row(s)
hive> ■
```

29. Now we similarly partition for course = Java and course = Python.

```
hive> load data local inpath "/home/cloudera/Desktop/Student.csv" into table student
      > partition(course="Python");
Loading data to table rjcstudent.student partition (course=Python)
Partition rjcstudent.student{course=Python} stats: [numFiles=1, numRows=0, totalSize=122, rawDataSize=0]
OK
Time taken: 1.687 seconds
hive> load data local inpath "/home/cloudera/Desktop/Student.csv" into table student
      > partition(course="Java");
Loading data to table rjcstudent.student partition (course=Java)
Partition rjcstudent.student{course=Java} stats: [numFiles=1, numRows=0, totalSize=122, rawDataSize=0]
OK
Time taken: 0.762 seconds
hive>
```

cloudera@quickstart:~/Desktop

30. Now go to browser refresh the page and select database as rjcstudent and click in

student.id	student.name	student.age	student.course
1	NULL	NULL	Hadoop
2	Renu	NULL	Hadoop
3	Shiv	NULL	Hadoop
4	Pratap	NULL	Hadoop

31. Now dropping the table student and creating the student table again normally (i.e. without partitioning).

drop table student;

0 SUCCESS.

```
create table student(ID int, Name string, Course string, Age int)
```

row format delimited

fields terminated by ','

tblproperties("skip.header.line.count"="1");

```
hive> create table student(ID int, Name string, Course string, Age int)
      > row format delimited
      > fields terminated by ','
      > tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.145 seconds
hive>
```

cloudera@quickstart:~/Desktop

32. Loading data in the student table from Student.csv file which present inside

/home/cloudera/Desktop directory.

load data local inpath "/home/cloudera/Desktop/Student.csv" into table student;

```
hive> load data local inpath "/home/cloudera/Desktop/Student.csv" into table student;
Loading data to table rjcstudent.student
Table rjcstudent.student stats: [numFiles=1, totalSize=122]
OK
Time taken: 0.482 seconds
hive>
```

Student.csv (~/Desktop) - gedit

select * from student;

```
hive> select * from student;
OK
1    Rahul  Hadoop  20
2    Shiv   Java     22
3    Pratap Python   23
4    Aakash SQL     26
5    Vinay  Java     27
6    Adesh  ReactJs 21
Time taken: 0.14 seconds, Fetched: 6 row(s)
hive>
```

33. Now creating employee.csv file.

gedit Employee.csv

cloudera-quickstart-vm-5.13.0-0-virtualbox (Running) - Oracle VM VirtualBox

File Machine View Input Devices Help
Applications Places System

```
cloudera@quickstart:~$ ls
cloudera-manager  Documents  enterprise-deployment.js
cm_api.py          Downloads  express-deployment.json
departments.java   dpt.java   kerberos
Desktop           eclipse   lib
cloudera@quickstart:~$ cd Desktop/
cloudera@quickstart:~/Desktop$ gedit Student.csv
cloudera@quickstart:~/Desktop$ gedit Employee.csv
cloudera@quickstart:~/Desktop$
```

Employee.csv (~/Desktop) - gedit

ID	Name	Dept	Yoj	Salary
1	Rahul	IT	2012	20000
2	Shiv	HR	2012	23000
3	Pratap	Sales	2013	25000
4	Aakash	HR	2014	22000

```
3    Pratap Python  23
4    Aakash SQL     26
5    Vinay  Java     27
6    Adesh  ReactJs 21
Time taken: 0.14 seconds, Fetched: 6 row(s)
hive>
```

cloudera@quickstart:~\$ [Hue - Editor - Mozilla... cloudera@quickstart... Employee.csv (~/Des...]

Querying :

34. Creating new database for performing querying operations.

create database hiveql;

```
hive> create database hiveql;
```

OK

Time taken: 0.082 seconds

```
hive> [cloudera@quickstart:~/Desktop]
```

cloudera@quickstart:~/Desktop

...

35. Using database hiveql and creating table employee inside the hiveql database.

create table employee(ID int, Name string, Department string, YOJ int, Salary float)

row format delimited

fields terminated by ','

tblproperties("skip.header.line.count"="1");

```
hive> create table employee(ID int, Name string, Department string, YOJ int, Salary float)
      > row format delimited
      > fields terminated by ','
      > tblproperties("skip.header.line.count"="1");
```

OK

Time taken: 0.113 seconds

```
hive> [Employee.csv (~/Desktop) - gedit]
```

...

Now we will see the schema of employee table using below command,

describe employee;

```
hive> describe employee;
OK
id          int
name        string
department  string
yoj         int
salary      float
Time taken: 0.184 seconds, Fetched: 5 row(s)
```

36. Loading the data into employee table from employee.csv file which we have created

earlier and it is present in /home/cloudera/Documents directory.

load data local inpath "/home/cloudera/Desktop/Employee.csv" into table employee;

```
hive> load data local inpath "/home/cloudera/Desktop/Employee.csv" into table employee;
Loading data to table rjcstudent.employee
Table rjcstudent.employee stats: [numFiles=1, totalSize=116]
OK
Time taken: 0.609 seconds
```

Displaying the table using below command,

select * from employee;

```
hive> select * from employee;
OK
1    Rahul   IT      2012    20000.0
2    Shiv    HR      2012    23000.0
3    Pratap  Sales   2013    25000.0
4    Aakash  HR      2014    22000.0
Time taken: 0.141 seconds, Fetched: 4 row(s)
```

```
hive> [Employee.csv (~/Desktop) - gedit]
```

...

Now we Perform some operations on this employee table.

37. We are printing or displaying the rows or records of employees whose salary is greater than and equal to 25000.

select * from employee where salary >=25000;

```
hive> select * from employee where salary >=25000;
Query ID = cloudera_20220326025858_4b2362ef-bd75-4923-9285-d373dbb34163
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1648262911818_0005, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648262911818_0005/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648262911818_0005
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-03-26 02:58:54,227 Stage-1 map = 0%, reduce = 0%
2022-03-26 02:59:16,852 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.44 sec
MapReduce Total cumulative CPU time: 4 seconds 440 msec
Ended Job = job_1648262911818_0005
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 4.44 sec HDFS Read: 4735 HDFS Write: 28 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 440 msec
OK
3 Pratap Sales 2013 25000.0
Time taken: 55.82 seconds, Fetched: 1 row(s)
hive> ■
```

No Items in Trash

38. Now we are printing or displaying the rows or records of employees whose salary is less than 25000.

select * from employee where salary <25000;

```
hive> select * from employee where salary <25000;
Query ID = cloudera_20220326030000_38afaa20-0321-413f-b867-18a4e55a7ab8
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1648262911818_0006, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648262911818_0006/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648262911818_0006
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-03-26 03:00:44,357 Stage-1 map = 0%, reduce = 0%
2022-03-26 03:01:09,379 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.62 sec
MapReduce Total cumulative CPU time: 4 seconds 620 msec
Ended Job = job_1648262911818_0006
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 4.62 sec HDFS Read: 4797 HDFS Write: 72 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 620 msec
OK
1 Rahul IT 2012 20000.0
2 Shiv HR 2012 23000.0
4 Aakash HR 2014 22000.0
Time taken: 58.944 seconds, Fetched: 3 row(s)
hive> ■
```

No Items in Trash

Aggregating

39. Arithmetic operations:

We are **adding** 2000 in existing salary and displaying specific columns using below command,

select ID, Name, salary + 2000 from employee;

```
hive> select ID, Name, salary + 2000 from employee;
Query ID = cloudera_20220326030303_332505b8-4680-4457-aaeb-ad9c82963858
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1648262911818_0007, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648262911818_0007/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648262911818_0007
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-03-26 03:52,983 Stage-1 map = 0%, reduce = 0%
2022-03-26 03:04:17,932 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.43 sec
MapReduce Total cumulative CPU time: 4 seconds 430 msec
Ended Job = job_1648262911818_0007
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 4.43 sec HDFS Read: 4881 HDFS Write: 65 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 430 msec
OK
1 Rahul 22000.0
2 Shiv 25000.0
3 Pratap 27000.0
4 Aakash 24000.0
Time taken: 64.708 seconds, Fetched: 4 row(s)
hive> ■
```

No Items in Trash

40. Displaying the **maximum salary** using below command,

select max(salary) from employee;

Name: Pavan Yadav

Practical 7

```

hive> select max(salary) from employee;
Query ID = cloudera_20220326030505_1f91277d-6abe-467d-8a66-c8e7e234dde1
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1648262911818_0008, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648262911818_0008/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648262911818_0008
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-03-26 03:06:17,722 Stage-1 map = 0%, reduce = 0%
2022-03-26 03:06:37,510 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.79 sec
2022-03-26 03:06:58,636 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.61 sec
MapReduce Total cumulative CPU time: 6 seconds 610 msec
Ended Job = job_1648262911818_0008
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.61 sec HDFS Read: 8055 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 610 msec
OK
25000.0
Time taken: 71.963 seconds, Fetched: 1 row(s)
hive> [cloudera@quickstart:~/Desktop]

```

Here we can see that 25000 is maximum salary.

41. Displaying the **minimum salary** using below command,

select min(salary) from employee;

```

hive> select min(salary) from employee;
Query ID = cloudera_20220326030707_6bbaa9ee-0051-4f2d-956a-197b44b53a46
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1648262911818_0009, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648262911818_0009/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648262911818_0009
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-03-26 03:08:30,966 Stage-1 map = 0%, reduce = 0%
2022-03-26 03:08:51,622 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.84 sec
2022-03-26 03:09:13,784 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.81 sec
MapReduce Total cumulative CPU time: 6 seconds 810 msec
Ended Job = job_1648262911818_0009
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.81 sec HDFS Read: 8055 HDFS Write: 8 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 810 msec
OK
20000.0
Time taken: 77.906 seconds, Fetched: 1 row(s)
hive> [No Items in Trash]

```

Here we can see that 20000 is minimum salary.

42. Now finding the **square root of salary** using below command,

select ID, Name, sqrt(salary) from employee;

```

hive> select ID, Name, sqrt(salary) from employee;
Query ID = cloudera_20220326031010_0793d885-79ca-4f9c-819d-f2d7598ed1f7
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1648262911818_0010, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648262911818_0010
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648262911818_0010
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-03-26 03:10:56,356 Stage-1 map = 0%, reduce = 0%
2022-03-26 03:11:17,978 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.65 sec
MapReduce Total cumulative CPU time: 4 seconds 650 msec
Ended Job = job_1648262911818_0010
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 4.65 sec HDFS Read: 4545 HDFS Write: 108 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 650 msec
OK
1      Rahul    141.4213562373095
2      Shiv     151.65750888103102
3      Pratap   158.11388300841898
4      Aakash   148.32396974191326
Time taken: 51.912 seconds, Fetched: 4 row(s)
hive> [cloudera@quickstart:~]

```

43. Now we are showing the name column in **upper case** using below command,

select ID, upper(name) from employee;

```

hive> select ID, upper(name) from employee;
Query ID = cloudera_20220326031212_8d0193e4-eea2-4fac-bdca-39b329e488b8
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1648262911818_0011, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648262911818_0011/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648262911818_0011
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-03-26 03:12:32,143 Stage-1 map = 0%, reduce = 0%
2022-03-26 03:12:53,889 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.75 sec
MapReduce Total cumulative CPU time: 4 seconds 750 msec
Ended Job = job_1648262911818_0011
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 4.75 sec HDFS Read: 4331 HDFS Write: 33 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 750 msec
OK
1      RAHUL
2      SHIV
3      PRATAP
4      AAKASH
Time taken: 52.859 seconds, Fetched: 4 row(s)
hive> [Hue - Editor - Mozilla Firefox]

```

44. Now we are showing the name column in **lower case** using below command,

select ID, lower(name) from employee;

```
hive> select ID, lower(name) from employee;
Query ID = cloudera_20220326031313_fbbd273b-decf-40ae-be8f-86447f24bd5e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1648262911818_0012, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648262911818_0012/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648262911818_0012
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2022-03-26 03:14:14,234 Stage-1 map = 0%, reduce = 0%
2022-03-26 03:14:39,884 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 4.53 sec
MapReduce Total cumulative CPU time: 4 seconds 530 msec
Ended Job = job_1648262911818_0012
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Cumulative CPU: 4.53 sec HDFS Read: 4331 HDFS Write: 33 SUCCESS
Total MapReduce CPU Time Spent: 4 seconds 530 msec
OK
1      rahul
2      shiv
3      pratap
4      aakash
Time taken: 53.377 seconds, Fetched: 4 row(s)
hive> ■
```

45. Creating another **employee2.csv** file with the same data as **employee.csv** but adding one more column as **Country** to it.

46. Creating a new table as **empgroup**.

create table empgroup(ID int, Name string, Department string, YOJ int, Salary float, Country string)

row format delimited

fields terminated by ','

tblproperties("skip.header.line.count"="1");

```
hive> create table empgroup(ID int, Name string, Department string, YOJ int, Salary float, Country string)
> row format delimited
> fields terminated by ','
> tblproperties("skip.header.line.count"="1");
OK
Time taken: 0.206 seconds
hive> ■
```

47. Loading the data into **empgroup** table from **employee2.csv** file which we have created and it is present in **/home/cloudera/Desktop** directory.

load data local inpath "/home/cloudera/Desktop/Employee2.csv" into table empgroup;

```
hive> load data local inpath "/home/cloudera/Desktop/Employee2.csv" into table empgroup;
Loading data to table rjcstudent.empgroup
Table rjcstudent.empgroup stats: [numFiles=1, totalSize=138]
OK
Time taken: 0.868 seconds
hive> ■
```

Displaying the table using below command,

Select * from empgroup;

```
hive> select * from empgroup;
OK
1      Rahul    IT      2012    20000.0 IND
2      Shiv     HR      2012    23000.0 UK
3      Pratap   Sales   2013    25000.0 UK
4      Aakash   HR      2014    22000.0 USA
Time taken: 0.129 seconds, Fetched: 4 row(s)
hive> ■
```

cloudera@quickstart:~/Desktop

48. Groupby clause

Now we display the total sum of salary of employees country wise using below command,

select Country, sum(Salary) from empgroup group by Country;

```
hive> select Country, sum(Salary) from empgroup group by Country;
Query ID = cloudera_20220326032121_506b17ce-b13e-4ae0-98c7-d4ff365af9ee
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1648262911818_0013, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648262911818_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648262911818_0013
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-03-26 03:22:03,266 Stage-1 map = 0%, reduce = 0%
2022-03-26 03:22:25,810 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.87 sec
2022-03-26 03:22:43,484 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 6.87 sec
MapReduce Total cumulative CPU time: 6 seconds 870 msec
Ended Job = job_1648262911818_0013
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 6.87 sec HDFS Read: 8629 HDFS Write: 35 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 870 msec
OK
IND      20000.0
UK       48000.0
USA     22000.0
Time taken: 69.464 seconds, Fetched: 3 row(s)
hive> ■
```

Groupby clause along with the having clause

Taking the total sum of salary countrywise using groupby clause and from that selecting or displaying those country whose **total sum of salary > 30000** using having clause.

select Country, sum(Salary) from empgroup group by Country having sum(Salary)>30000;

```
hive> select Country, sum(Salary) from empgroup group by Country having sum(Salary)>30000;
Query ID = cloudera_20220326032525_73a66c90-d2d6-4322-8a24-ab2fee5fe74e
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1648262911818_0015, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648262911818_0015/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648262911818_0015
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-03-26 03:25:43,987 Stage-1 map = 0%, reduce = 0%
2022-03-26 03:26:03,370 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.31 sec
2022-03-26 03:26:27,298 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 8.62 sec
MapReduce Total cumulative CPU time: 8 seconds 620 msec
Ended Job = job_1648262911818_0015
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 8.62 sec HDFS Read: 9138 HDFS Write: 11 SUCCESS
Total MapReduce CPU Time Spent: 8 seconds 620 msec
OK
UK      48000.0
Time taken: 69.319 seconds, Fetched: 1 row(s)
hive> ■
```

Sorting : Order by

49. Now we are displaying the table by sorting the salary in **descending order**.

select * from empgroup order by Salary desc;

```
hive> select * from empgroup order by Salary desc;
Query ID = cloudera_20220326032828_d75db91e-d98a-4a2a-b3ac-4f8ddad78660
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks determined at compile time: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1648262911818_0016, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648262911818_0016/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648262911818_0016
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-03-26 03:28:41,036 Stage-1 map = 0%, reduce = 0%
2022-03-26 03:28:59,643 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.3 sec
2022-03-26 03:29:23,519 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 7.09 sec
MapReduce Total cumulative CPU time: 7 seconds 90 msec
Ended Job = job_1648262911818_0016
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1 Reduce: 1 Cumulative CPU: 7.09 sec HDFS Read: 8624 HDFS Write: 114 SUCCESS
Total MapReduce CPU Time Spent: 7 seconds 90 msec
OK
3    Pratap   Sales   2013   25000.0 UK
2    Shiv     HR      2012   23000.0 UK
4    Aakash   HR      2014   22000.0 USA
1    Rahul    IT      2012   20000.0 IND
Time taken: 70.666 seconds, Fetched: 4 row(s)
hive> ■
```

We can see that number of reducers: 1 for the order by.

50. Now we are displaying the table by **sorting** the salary in descending order.

select * from empgroup sort by Salary desc;

```
hive> select * from empgroup sort by Salary desc;
Query ID = cloudera_20220326033030_b0aa6056-13eb-48bb-aa8c-c1e56c95caec
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Starting Job = job_1648262911818_0017, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1648262911818_0017/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1648262911818_0017
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2022-03-26 03:30:55,515 Stage-1 map = 0%,  reduce = 0%
2022-03-26 03:31:20,041 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 2.83 sec
2022-03-26 03:31:43,167 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 6.78 sec
MapReduce Total cumulative CPU time: 6 seconds 780 msec
Ended Job = job_1648262911818_0017
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1  Cumulative CPU: 6.78 sec  HDFS Read: 8624 HDFS Write: 114 SUCCESS
Total MapReduce CPU Time Spent: 6 seconds 780 msec
OK
3      Pratap  Sales   2013    25000.0 UK
2      Shiv    HR      2012    23000.0 UK
4      Aakash  HR      2014    22000.0 USA
1      Rahul   IT      2012    20000.0 IND
Time taken: 79.267 seconds, Fetched: 4 row(s)
hive> ■
```

Now we can see the similar result as we got from order by and sort by so what is the difference between the two is that it depends on number of reducers in order by we got number of reducers is 1 and by using sort by here is also we got number of reducers is 1 so the difference between the two is that Order by will guarantee the total order in the output whereas sort by will only guarantee the ordering of the rows within the reducer.

Order by gives us completely sorted result whereas sort by give us partially sorted result.

quit

```
hive> quit
> ;
WARN: The method class org.apache.commons.logging.impl.SLF4JLogFactory#release() was invoked.
WARN: Please see http://www.slf4j.org/codes.html#release for an explanation.
[cloudera@quickstart ~]$ ■
```

cloudera@quickstart:~/Desktop

The screenshot shows a terminal window with the title bar 'cloudera@quickstart:~'. Inside the window, the command 'quit' is typed and executed. The output shows two 'WARN' messages about the SLF4J Log Factory release method. The prompt '[cloudera@quickstart ~]\$' is visible at the bottom, followed by a red square cursor character.

Aim: To implement Word Count problem using Pig

What is Apache Pig

Apache Pig is a high-level data flow platform for executing MapReduce programs of Hadoop. The language used for Pig is Pig Latin.

The Pig scripts get internally converted to Map Reduce jobs and get executed on data stored in HDFS. Apart from that, Pig can also execute its job in Apache Tez or Apache Spark.

Pig can handle any type of data, i.e., structured, semi-structured or unstructured and stores the corresponding results into Hadoop Data File System. Every task which can be achieved using PIG can also be achieved using java used in MapReduce.

Features of Apache Pig

The various uses of Pig technology.

1) Ease of programming

Writing complex java programs for map reduce is quite tough for non-programmers. Pig makes this process easy. In the Pig, the queries are converted to MapReduce internally.

2) Optimization opportunities

It is how tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.

3) Extensibility

A user-defined function is written in which the user can write their logic to execute over the data set.

4) Flexible

It can easily handle structured as well as unstructured data.

5) In-built operators

It contains various type of operators such as sort, filter and joins.

Advantages of Apache Pig

- Less code - The Pig consumes less line of code to perform any operation.
- Reusability - The Pig code is flexible enough to reuse again.
- Nested data types - The Pig provides a useful concept of nested data types like tuple, bag, and

Before the perform the practical first perform 3 steps of the given following

1. sudo /home/cloudera/cloudera-manager --force --express

```

cloudera@quickstart-vm-5-13-0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Fri Mar 25, 1:15 AM cloudera@quickstart:~ cloudera@quickstart:~

[QuickStart] Waiting for Cloudera Manager API...
[QuickStart] Starting Cloudera Manager agent...
[QuickStart] Configuring deployment...
Submitted jobs: 41
[QuickStart] Deploying client configuration...
Submitted jobs: 42
[QuickStart] Starting Cloudera Management Service...
Submitted jobs: 58
[QuickStart] Enabling Cloudera Manager daemons on boot...

Success! You can now log into Cloudera Manager from the QuickStart VM's browser:
http://quickstart.cloudera:7180

Username: cloudera
Password: cloudera

[cloudera@quickstart ~]$ sudo -u hdfs hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

22/03/25 01:04:22 WARN ipc.Client: Failed to connect to server: quickstart.cloudera/10.0.2.15:8020; try once and fail.
java.net.ConnectException: Connection refused
at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
at sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:739)
at org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout.java:206)
at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:538)
at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:494)
at org.apache.hadoop.ipc.Client$Connection.setupConnection(Client.java:648)
at org.apache.hadoop.ipc.Client$Connection.access$3000(Client.java:744)
at org.apache.hadoop.ipc.Client$Connection.access$3000(Client.java:399)

cloudera@quickstart:~$ Home - Cloudera Man...

```

2. Start all services

The screenshot shows the Cloudera Manager interface. On the left, there's a sidebar with 'Cloudera Quickstart' and various service icons: Hosts (1), HBase (1), HDFS (2), Hive (2), Hue (1), Impala (1), and Key-Value Store (1). The main area has four charts: Cluster CPU, Cluster Disk IO, Cluster Network IO, and HDFS IO. At the top, there are tabs for Status, Configuration, and All Recent Commands.

This screenshot shows the 'Restart Command' dialog in Cloudera Manager. It lists several steps: 'Start All Services' (Completed), 'Format HDFS' (In Progress), 'Format HDFS' (In Progress), 'Format HDFS' (In Progress), and 'Format HDFS' (In Progress). The 'Completed' step took 7m.

3. sudo -u hdfs hadoop dfsadmin -safemode leave

```

cloudera@quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~ Fri Mar 25, 1:16 AM cloudera
File Edit View Search Terminal Help
at org.apache.hadoop.ipc.ClientsConnection.setupConnection(Client.java:648)
at org.apache.hadoop.ipc.ClientsConnection.setupIOstreams(Client.java:744)
at org.apache.hadoop.ipc.ClientsConnection.access$3000(Client.java:386)
at org.apache.hadoop.ipc.ClientsConnection.getConnection(Client.java:1557)
at org.apache.hadoop.ipc.Client.call(Client.java:1488)
at org.apache.hadoop.ipc.Client.call(Client.java:1441)
at org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngine.java:230)
at com.sun.proxy.$Proxy16.setSafeMode(Unknown Source)
at org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.setSafeMode(ClientNamenodeProtocolTranslatorPB.java:681)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:606)
at org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryInvocationHandler.java:288)
at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:104)
at com.sun.proxy.$Proxy17.setSafeMode(Unknown Source)
at org.apache.hadoop.hdfs.DFSClient.setSafeMode(DFSClient.java:2648)
at org.apache.hadoop.hdfs.DistributedFileSystem.setSafeMode(DistributedFileSystem.java:1182)
at org.apache.hadoop.hdfs.DistributedFileSystem.setSafeMode(DistributedFileSystem.java:1166)
at org.apache.hadoop.hdfs.tools.DFSAdmin.setSafeMode(DFSAdmin.java:576)
at org.apache.hadoop.hdfs.tools.DFSAdmin.run(DFSAdmin.java:1856)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:84)
at org.apache.hadoop.hdfs.tools.DFSAdmin.main(DFSAdmin.java:2032)

safeMode: Call From quickstart.cloudera:18.0.2.15 to quickstart.cloudera:8020 failed on connection exception: java.net.ConnectException: Connection refused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
[cloudera@quickstart ~]$ sudo -u hdfs hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Safe mode is OFF
[cloudera@quickstart ~]$ 

```

1. Open the browser. And then open Hue and login.

This screenshot shows the Cloudera Manager interface again. The sidebar shows the same service counts as before. The main area features four charts: Cluster CPU, Cluster Disk IO, Cluster Network IO, and HDFS IO. The top navigation bar includes tabs for Status, Configuration, and All Recent Commands.

Name: Pavan Yadav

Practical 8

Hue - Welcome to Hue - Mozilla Firefox

Fri Mar 25, 6:34 AM cloudera

Home - Cloudera Ma... Hue - Welcome to Hue

quickstart.cloudera:8888/accounts/login/?next=/hue

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

HUE

Query. Explore. Repeat.

cloudera

Sign In

Hue and the Hue logo are trademarks of Cloudera, Inc.

2. Now open the directory /user/cloudera

Type /hue/filebrowser

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Applications Places System

Fri Mar 25, 6:39 AM cloudera

Home - Cloudera Ma... Hue - File Browser

quickstart.cloudera:8888/hue/filebrowser/view=/user/cloudera

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

HUE

Query

File Browser

Search for file name

Actions Move to trash

Upload New

File Directory

Home / user / cloudera

Name	Size	User	Group	Permissions	Date
hdfs		supergroup	supergroup	drwxr-xr-x	March 17, 2022 07:20 AM
.		cloudera	cloudera	drwxr-xr-x	March 17, 2022 07:18 AM
.Trash		cloudera	cloudera	drwx-----	March 18, 2022 08:01 AM

Show 45 of 1 items

Page 1 of 1

3. Now we are creating the directory as Training inside /user/cloudera

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox

File Machine View Input Devices Help

Applications Places System

Fri Mar 25, 2:20 AM cloudera

Home - Cloudera Ma... Hue - File Browser

quickstart.cloudera:8888/hue/filebrowser/view=/user/cloudera

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

HUE

Query

File Browser

Search for file name

Actions Move to trash

Upload New

File Directory

Home / user / cloudera

Name	Size	User	Group	Permissions	Date
hdfs		supergroup	supergroup	drwxr-xr-x	October 23, 2017 09:17 AM
.		cloudera	cloudera	drwxr-xr-x	October 23, 2017 09:14 AM

Show 45 of 0 items

Page 1 of 1

Hue - File Browser - Mozilla Firefox

Create Directory

Directory Name: Training

Name	Size	User	Group	Permissions	Date
hdfs		supergroup	supergroup	drwxr-xr-x	March 17, 2022 07:20 AM
.		cloudera	cloudera	drwxr-xr-x	March 17, 2022 07:18 AM
..		cloudera	cloudera	drwxr-xr-x	March 18, 2022 08:01 AM
.Trash		cloudera	cloudera	drwxr-xr-x	March 18, 2022 08:01 AM
Training		cloudera	cloudera	drwxr-xr-x	March 25, 2022 06:41 AM

Hue - File Browser - Mozilla Firefox

Search data and saved documents...

Name	Size	User	Group	Permissions	Date
hdfs		supergroup	supergroup	drwxr-xr-x	March 17, 2022 07:20 AM
.		cloudera	cloudera	drwxr-xr-x	March 25, 2022 06:41 AM
..		cloudera	cloudera	drwxr-xr-x	March 18, 2022 08:01 AM
.Trash		cloudera	cloudera	drwxr-xr-x	March 18, 2022 08:01 AM
Training		cloudera	cloudera	drwxr-xr-x	March 25, 2022 06:41 AM

4. After creating Training directory now creating the **pig** directory inside /user/cloudera/Training

Hue - File Browser - Mozilla Firefox

File Browser

Home / user / cloudera / Training

Name	Size	User	Group	Permissions	Date
hdfs		cloudera	cloudera	drwxr-xr-x	March 25, 2022 06:41 AM
pig		cloudera	cloudera	drwxr-xr-x	March 25, 2022 06:41 AM
Training		cloudera	cloudera	drwxr-xr-x	March 25, 2022 06:41 AM

The screenshot shows the Hue File Browser interface. A modal dialog box titled "Create Directory" is open, with the "Directory Name" field set to "pig". In the background, the main file browser view shows a list of items. Two entries are visible: one named "_" and another named "pig". Both entries have "cloudera" listed under "User" and "Group", and "drwxr-xr-x" under "Permissions". The date for both entries is "March 25, 2022 06:41 AM". The status bar at the bottom left indicates the user is at the command prompt "cloudera@quickstart:~".

pig directory has been created inside /user/cloudera/Training

The screenshot shows the Hue File Browser interface. The current path is displayed as "/user/cloudera/Training". Inside this directory, there is a single entry named "pig", which is highlighted with a red border. The status bar at the bottom left indicates the user is at the command prompt "cloudera@quickstart:~".

5. Creating input.txt file inside /usr/cloudera/training/pig directory

The screenshot shows the Hue File Browser interface. The current path is displayed as "/user/cloudera/Training/pig". A context menu is open over the "pig" directory, with the "New" option selected. Under "New", the "File" option is highlighted with a red border. The status bar at the bottom left indicates the user is at the command prompt "cloudera@quickstart:~".

The screenshot shows the Cloudera Manager Hue - File Browser interface. A modal dialog titled 'Create File' is open, with the 'File Name' field containing 'input.txt'. In the background, the main file browser pane displays a list of files with the following details:

Name	Size	User	Group	Permissions	Date
input.txt		cloudera	cloudera	drwxr-xr-x	March 25, 2022 06:43 AM
input2.txt		cloudera	cloudera	drwxr-xr-x	March 25, 2022 06:43 AM

Note: If getting this type of error than start all services again

The screenshot shows the same 'Create File' dialog as the previous one, but now an error message is visible in the top right corner: 'IOException: Failed to find datanode, suggest to check cluster health. excludeDatanodes=null (error 403)'. The main file browser pane shows the same list of files as before.

The screenshot shows the Cloudera Manager Home page. On the left, the 'Cloudera QuickStart' interface is visible, with the 'Hosts' section showing 'Start' and 'Stop' buttons highlighted with red boxes. On the right, there are four charts: 'Cluster CPU' (CPU usage), 'Cluster Disk IO' (disk I/O rates), 'Cluster Network IO' (network traffic), and 'HDFS IO' (HDFS I/O). The status bar at the bottom indicates 'quickstart:cloudera:7180/cmft/home# cloudera@quickstart:~\$'.

Refresh the Page

The screenshot shows the Cloudera Manager Hue - File Browser after refreshing the page. The 'input.txt' entry from the previous list is no longer present, leaving only 'input2.txt' in the list.

The screenshot shows the Hue File Browser interface. The main view displays a list of files in the directory `/user/cloudera/Training/pig`. Two files are listed: `cloudera` and `cloudera`. In the top right corner of the browser window, there is a modal dialog titled "Create File". Inside the modal, there is a "File Name" input field containing the value "input.txt". Below the input field are "Cancel" and "Create" buttons. The "Create" button is highlighted with a blue background.

6. Adding some contents to this input.txt file.

The screenshot shows the Hue File Browser interface. The main view displays a list of files in the directory `/user/cloudera/Training/pig`. One file, `input.txt`, is selected and highlighted with a red border. The file details are shown in a table below the list:

Name	Size	User	Group	Permissions	Date
<code>input.txt</code>	0 bytes	cloudera	cloudera	-rw-r--r--	March 25, 2022 07:17 AM

Click on **Edit file** option

The screenshot shows the Hue File Browser interface. The URL in the address bar is `quickstart.cloudera:8888/hue/filebrowser/view=/user/cloudera/Training/pig/input.txt`. The left sidebar shows 'Sources' with 'Hive' selected. The main area displays the contents of 'input.txt' which is currently empty. There are buttons for 'Edit file', 'Download', 'View file location', and 'Refresh'. The status bar at the bottom shows 'cloudera@quickstart:~'.

Save the input.txt file

The screenshot shows the 'Edit file' dialog for 'input.txt'. The content of the file is:

```
We are studying Big Data Technology
We have covered the Hadoop ecosystem
We have executed Wordcount using MapReduce
We have also implemented CRUD operation using MongoDB
```

At the bottom of the dialog are 'Save' and 'Save as' buttons. The status bar at the bottom shows 'cloudera@quickstart:~'.

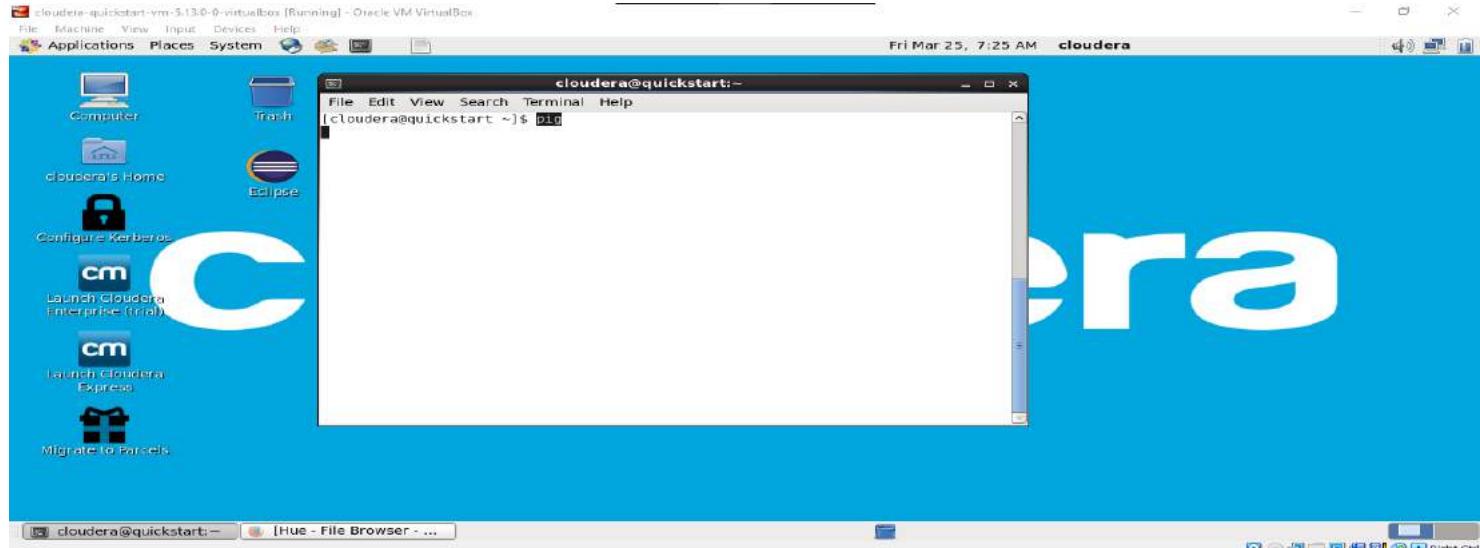
The screenshot shows the Hue File Browser interface again. The URL in the address bar is the same as before. The left sidebar shows 'Sources' with 'Hive' selected. The main area now displays the updated contents of 'input.txt'.

```
We are studying Big Data Technology
We have covered the Hadoop ecosystem
We have executed Wordcount using MapReduce
We have also implemented CRUD operation using MongoDB
```

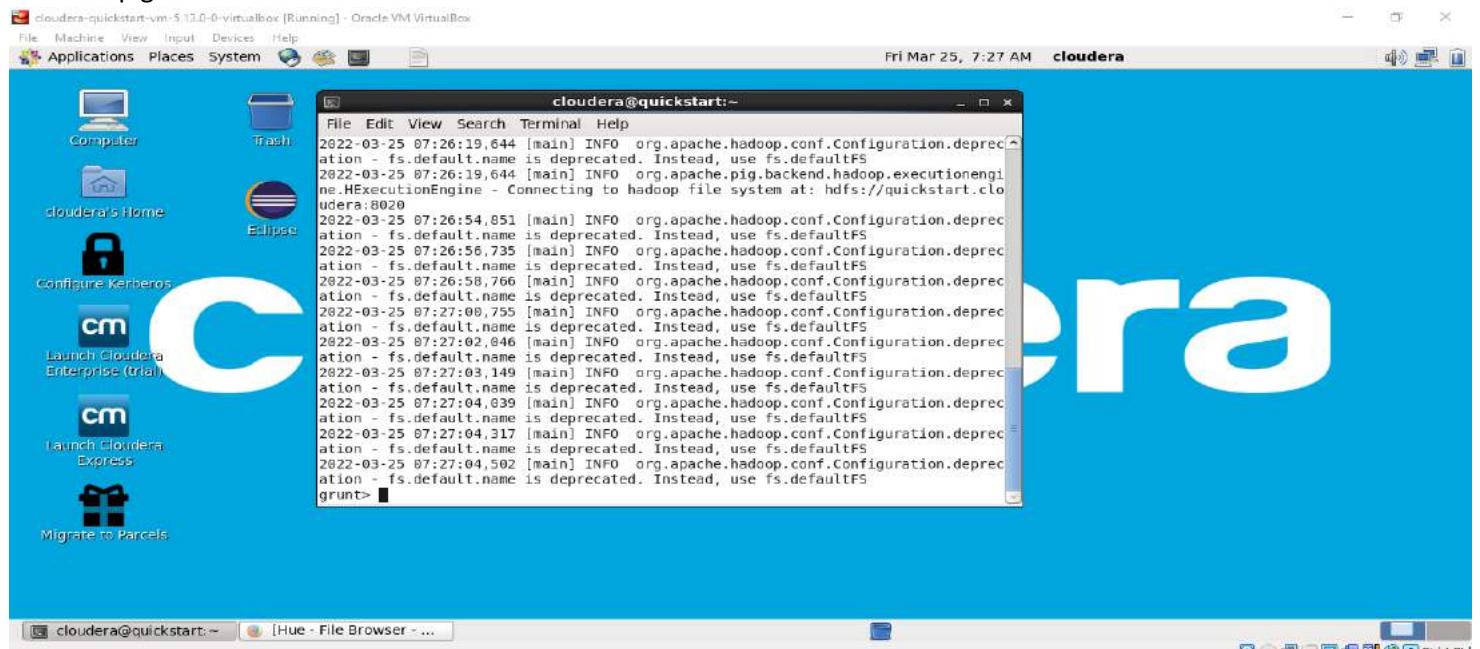
The status bar at the bottom shows 'cloudera@quickstart:~'.

7. Now Open the terminal. And start **Pig** by typing pig on terminal.

pig



Now the pig started



8. Now we have to load that input file where ever it is stored. By typing the command

input1 = LOAD '/user/cloudera/Training/pig/input.txt' AS (f1:chararray);



9. Now we are dumping the data. It will done the mapreduce task.

DUMP input1

A screenshot of a Linux desktop environment. On the left, there's a vertical dock with icons for Cloudera's Home, Eclipse, Configure Kerberos, cm (Launch Cloudera Enterprise (trial)), Launch Cloudera Express, and Migrate to Parcels. The main area shows a terminal window with the following log output from Apache Pig:

A screenshot of a Cloudera Quickstart VM desktop environment. The desktop has a blue theme with icons for various applications like Computer, Trash, and Eclipse. A large white 'C' logo is in the center. A terminal window titled 'cloudera@quickstart:' is open, displaying Apache Pig logs. The logs show the execution of a MapReduce job, including counts of records and bytes written, and details about the job DAG and input paths. The terminal ends with a 'grunt>' prompt.

10. wordsInEachLine = FOREACH input1 GENERATE flatten(TOKENIZE(f1)) as word;

```
2022-03-25 07:37:17,932 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
2022-03-25 07:37:17,952 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-25 07:37:17,961 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set... will not generate code.
2022-03-25 07:37:18,099 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-03-25 07:37:18,101 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
(we are studying Big Data Technology)
(we have covered the Hadoop ecosystem)
(we have executed Wordcount using MapReduce)
(we have also implemented CRUD operations using MongoDB)
grunts> wordsInEachLine = FOREACH Input1 GENERATE Flatten(TOKENIZE(f1)) as word;
grunts>
```

11. DUMP wordsInEachLine:

```
Configure Kerberos  
cm Launch Cloudera Enterprise (trial)  
cm Launch Cloudera Express  
Migrate to Parcels  
  
grunt> wordsInEachLine = FOREACH input1 GENERATE flatten(TOKENIZE(f1)) as word;  
grunt> DUMP wordsInEachLine;  
2022-03-25 07:41:25,391 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used  
in the script: UNKNOWN  
2022-03-25 07:41:25,408 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer  
- {RULES_ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParal  
lelSetter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach,  
NewPartitionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastIns  
ter], RULES_DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]}  
2022-03-25 07:41:25,490 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.  
MRCompiler - File concatenation threshold: 100 optimistic? false  
2022-03-25 07:41:25,507 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.  
MultiQueryOptimizer - MR plan size before optimization: 1  
2022-03-25 07:41:25,507 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.  
MultiQueryOptimizer - MR plan size after optimization: 1  
2022-03-25 07:41:25,762 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to Resourc  
eManager at quickstart.cloudera/10.0.2.15:8032  
2022-03-25 07:41:25,825 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settin  
gs are added to the job  
2022-03-25 07:41:26,032 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.  
JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
```

```

cloudera@quickstart:~$ input paths to process : 1
2022-03-25 07:44:23,576 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil
- Total input paths to process : 1
(we)
(are)
(studying)
(Big)
(Data)
(Technology)
(we)
(have)
(covered)
(the)
(Hadoop)
(ecosystem)
(We)
(have)
(executed)
(Wordcount)
(using)
(MapReduce)
(We)
(have)
(also)
(implemented)
(CRUD)
(operation)
(using)
(MongoDB)
grunt>

```

12. Now grouping the words present in each line.

groupedWords = group wordsInEachLine by word;

```

cloudera@quickstart:~$ input paths to process : 1
2022-03-25 07:44:23,576 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil
- Total input paths to process : 1
(we)
(are)
(studying)
(Big)
(Data)
(Technology)
(we)
(have)
(covered)
(the)
(Hadoop)
(ecosystem)
(We)
(have)
(executed)
(Wordcount)
(using)
(MapReduce)
(We)
(have)
(also)
(implemented)
(CRUD)
(operation)
(using)
(MongoDB)
grunt> groupedWords = group wordsInEachLine by word;
grunt>

```

13. DUMP groupedWords;

```

cloudera@quickstart:~$ input paths to process : 1
2022-03-25 07:44:23,576 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil
- Total input paths to process : 1
(we)
(are)
(studying)
(Big)
(Data)
(Technology)
(we)
(have)
(covered)
(the)
(Hadoop)
(ecosystem)
(We)
(have)
(executed)
(Wordcount)
(using)
(MapReduce)
(We)
(have)
(also)
(implemented)
(CRUD)
(operation)
(using)
(MongoDB)
grunt> groupedWords = group wordsInEachLine by word;
grunt> DUMP groupedWords;
2022-03-25 07:47:36,142 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in
the script: GROUP_BY
2022-03-25 07:47:36,146 [main] INFO org.apache.pig.newplan.logical.optimizer.LogicalPlanOptimizer -
#RULES ENABLED=[AddForEach, ColumnMapKeyPrune, DuplicateForEachColumnRewrite, GroupByConstParallelSet
ter, ImplicitSplitInserter, LimitOptimizer, LoadTypeCastInserter, MergeFilter, MergeForEach, NewPartit
ionFilterOptimizer, PushDownForEachFlatten, PushUpFilter, SplitFilter, StreamTypeCastInserter], RULE
S DISABLED=[FilterLogicExpressionSimplifier, PartitionFilterOptimizer]
2022-03-25 07:47:36,275 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRC
ompiler - File concatenation threshold: 100 optimistic? false
2022-03-25 07:47:36,312 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Mul
tiQueryOptimizer - MR plan size before optimization: 1
2022-03-25 07:47:36,312 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Mul
tiQueryOptimizer - MR plan size after optimization: 1
2022-03-25 07:47:36,448 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceMa
nager at quickstart.cloudera/10.0.2.15:8032
2022-03-25 07:47:36,452 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings
are added to the job
2022-03-25 07:47:36,508 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Job
ControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2022-03-25 07:47:36,502 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Job
ControlCompiler - Reduce phase detected, estimating # of required reducers.
2022-03-25 07:47:36,511 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Job
ControlCompiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLay
er.InputSizeReducerEstimator
2022-03-25 07:47:36,548 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Inp
utSizeReducerEstimator - BytesPerReducer=1000000000 maxReducers=999 totalInputFileSize=169
2022-03-25 07:47:36,548 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.Job
ControlCompiler - Setting Parallelism to 1

```

```

File Edit View Search Terminal Help
ReduceLauncher - Success!
2022-03-25 07:52:04,569 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-25 07:52:04,472 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set, will not generate code.
2022-03-25 07:52:04,569 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-03-25 07:52:04,569 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
{((we),(we),(we),(we))}
{Big,(1)}
{are,(are),(are),(are)}
{(CRUD,{(CRUD)})
{Data,{(Data)})
{the,(the),(the),(the)}
{(have,(have),(have),(have))}
{(using,(using),(using))}
{(MongoDB,{MongoDB})}
{(covered,{(covered)})
{(executed,{(executed)})
{(ready,{(ready)})
{(MapReduce,{(MapReduce)})
{(Wordcount,{(Wordcount)})
{(ecosystem,(ecosystem)
{(operation,{(operation)})
{(Technology,{(Technology)})
{(implemented,{(implemented)})
grants>

```

14. Now we count those words. For each group we count words in each line.

countedWords = FOREACH groupedWords GENERATE group, COUNT(wordsInEachLine);

```

Migrate to Parcels: <line 4, column 4>: invalid Field projection. Projected Field [GROUP] does not exist in schema: group :chararray,wordsInEachLine:bag{:tuple(word:chararray)}.
Details at logfile: /home/cloudera/pig_1648218365125.log
grunt> countedWords = FOREACH groupedWords GENERATE group, COUNT(wordsInEachLine);
grunt>

```

15. DUMP countedWords;

Now the Final Output we are getting as word count for every word.

```

File Edit View Search Terminal Help
Line 4, column 4>: invalid Field projection. Projected Field [GROUP] does not exist in schema: group :chararray,wordsInEachLine:bag{:tuple(word:chararray)}.
Details at logfile: /home/cloudera/pig_1648218365125.log
grunt> DUMP countedWords
2022-03-25 08:01:19,431 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in this script: GROUP
2022-03-25 08:01:19,438 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer - Pig features used in this script: GROUP
2022-03-25 08:01:19,524 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer - MR job identifier: quickstart.cloudera/18-60-15-002
2022-03-25 08:01:19,524 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer - MR job identifier: quickstart.cloudera/18-60-15-002
2022-03-25 08:01:19,684 [main] INFO org.apache.hadoop.yarn.client.RMProxy - Connecting to ResourceManager at quickstart.cloudera/18-60-15-002
2022-03-25 08:01:19,755 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2022-03-25 08:01:19,755 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer - Job contrl compiler: Using preprend job.reduce.markreset.buffer.percent is not set, set to default 0.3
2022-03-25 08:01:19,757 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer - Job control compiler: Using preprend job.reduce.markreset.buffer.percent is not set, set to default 0.3
2022-03-25 08:01:19,767 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer - Job Control Compiler - Using reducer estimator: org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.

```

```

File Edit View Search Terminal Help
ReduceLauncher - Success!
2022-03-25 08:11:27,868 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - fs.default.name is deprecated. Instead, use fs.defaultFS
2022-03-25 08:11:27,905 [main] INFO org.apache.pig.data.SchemaTupleBackend - Key [pig.schematuple] was not set, will not generate code.
2022-03-25 08:11:27,905 [main] INFO org.apache.hadoop.mapreduce.lib.input.FileInputFormat - Total input paths to process : 1
2022-03-25 08:11:27,905 [main] INFO org.apache.pig.backend.hadoop.executionengine.util.MapRedUtil - Total input paths to process : 1
{((we),(we),(we),(we))}
{Big,(1)}
{are,(are),(are),(are)}
{(CRUD,{(CRUD)})
{Data,{(Data)})
{the,(the),(the),(the)}
{(have,(have),(have),(have))}
{(using,(using),(using))}
{(MongoDB,{MongoDB})}
{(covered,{(covered)})
{(executed,{(executed)})
{(ready,{(ready)})
{(MapReduce,{(MapReduce)})
{(Wordcount,{(Wordcount)})
{(ecosystem,(ecosystem)
{(operation,{(operation)})
{(Technology,{(Technology)})
{(implemented,{(implemented)})
grants>

```

As we can see from above image the Word "We" start with capital W occurred four times, word "Big" occurred once, and so on.

Now Exit from the grunt shell using quit command.

16. quit

```

Migrate to Parcels: (Technology,1)
(implemented,1)
grunt> quit
[cloudera@quickstart ~]$ 

```

Aim: Demonstrate the use of Sqoop tool to transfer data between Hadoop & relational database servers

Sqoop:

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

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

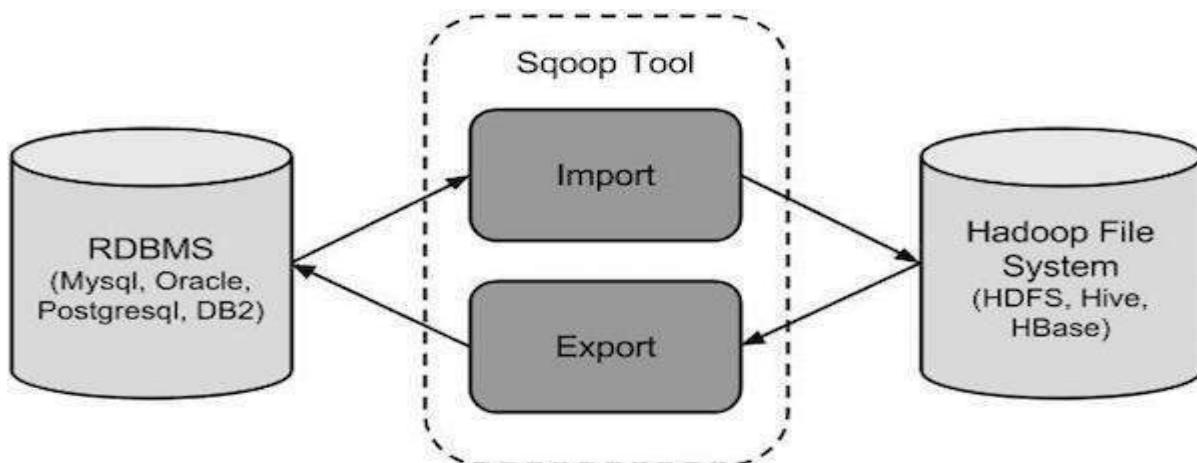
Sqoop – “SQL to Hadoop and Hadoop to SQL”

Sqoop is a tool designed to transfer data between Hadoop and external datastores such as relational databases and enterprise data warehouses.

It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases.

It imports data from external datastores into HDFS, Hive, and HBase.

Working of Sqoop:



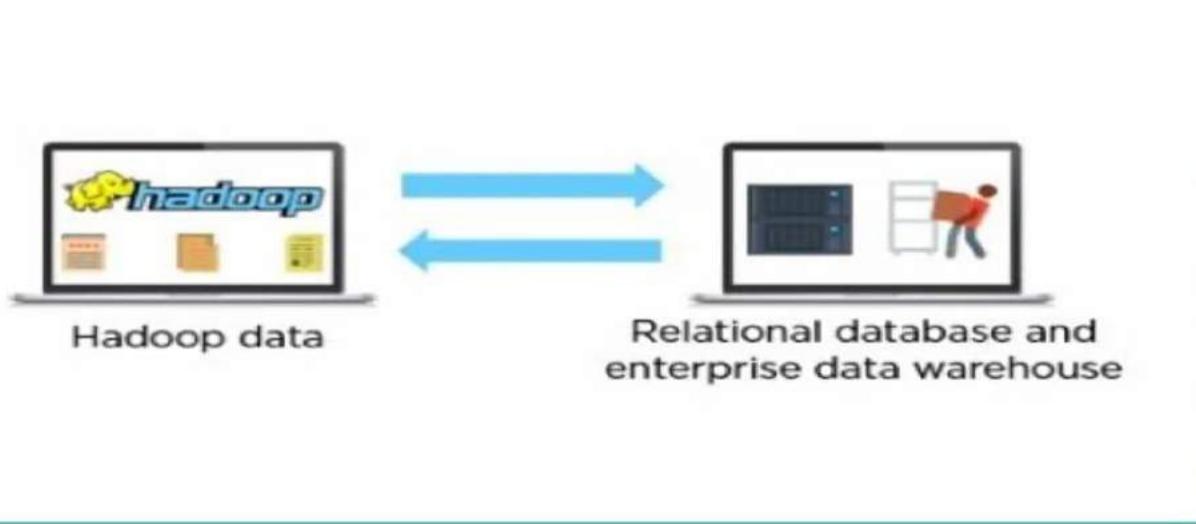
There are mainly two functions associated with Apache Sqoop tool which are Sqoop Import and Sqoop export.

1. Sqoop Import

Sqoop Import This is an important function which executes the task of data importing from external sources (RDBMS) to HDFS. In HDFS, each row of a table is considered as a record. The entire records are stored in a text format in the text files or as binary data in Sequence files.

2. Sqoop Export

Sqoop Export This function performs bulk data exportation tasks from the HDFS to RDBMS. Once the modifications are done to the imported records you will get a result set and the next process is to send back the data to the relational database (RDBMS). Sqoop export function reads a group of delimited text files from HDFS in parallel, divides the files into records, and stores them as new rows in a targeted database table.

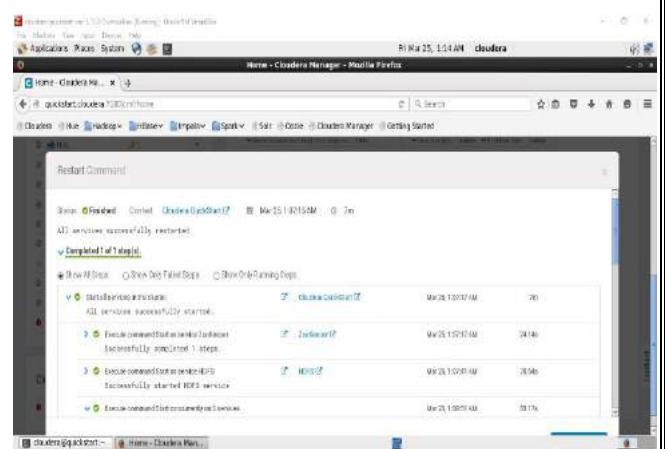
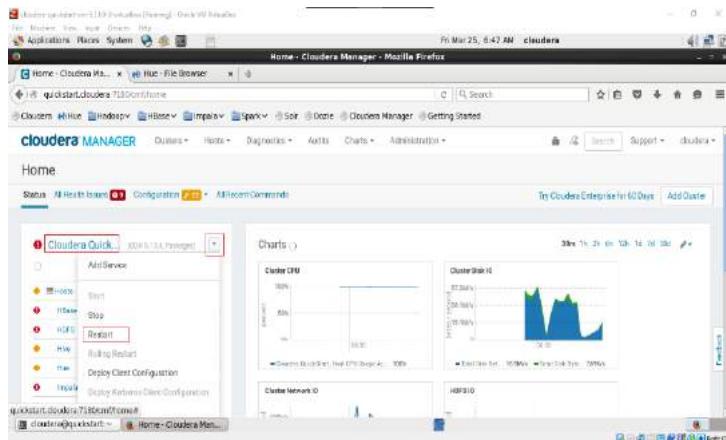


Before performing the practical first perform 3 steps of the given following

1. sudo /home/cloudera/cloudera-manager --force --express

```
cloudera@quickstart-vm-5:130-0-virtualbox:~$ sudo /home/cloudera/cloudera-manager --force --express
File Machine View Input Devices Help
Applications Places System
Fri Mar 25, 1:15 AM cloudera
File Edit View Search Terminal Help
[QuickStart] Waiting for Cloudera Manager API...
[QuickStart] Starting Cloudera Manager agent...
[QuickStart] Configuring deployment...
Submitted jobs: 4
[QuickStart] Deploying client configuration...
Submitted jobs: 42
[QuickStart] Starting Cloudera Management Service...
Submitted jobs: 58
[QuickStart] Enabling Cloudera Manager daemons on boot...
Success! You can now log into Cloudera Manager from the QuickStart VM's browser:
http://quickstart.cloudera:7180
Username: cloudera
Password: cloudera
[cloudera@quickstart ~]$ sudo -u hdfs hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.
22/03/25 01:04:22 WARN ipc.Client: Failed to connect to server: quickstart.cloudera/10.0.2.15:8820: try once and fail.
java.net.ConnectException: Connection refused
at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
at sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:739)
at org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout.java:206)
at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:528)
at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:494)
at org.apache.hadoop.ipc.Client$Connection.setupConnection(Client.java:648)
at org.apache.hadoop.ipc.Client$Connection.setupIOstreams(Client.java:744)
at org.apache.hadoop.ipc.Client$Connection.access$3000(Client.java:390)
```

2. Start all services



3. sudo -u hadoop dfsadmin -safemode leave

```

cloudera@quickstart:~$ sudo -u hadoop dfsadmin -safemode leave
[sudo] password for cloudera: 
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Safe mode is OFF
[cloudera@quickstart ~]$ 

```

Steps: Demonstrate the use of Sqoop tool to transfer data between Hadoop & relational database servers

- Starting the mysql by giving username as root and password as cloudera.

mysql -u root -pcloudera

```

cloudera@quickstart:~$ mysql -u root -pcloudera
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 432
Server version: 5.1.73 Source distribution

Copyright (c) 2009, 2013, Oracle and/or its affiliates. All rights reserved.

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> 

```

- Now using below command it will displaying or give the list of databases which are already present or exist in mysql.

show databases;

```

cloudera@quickstart:~$ show databases;
+-----+
| Database |
+-----+
| information_schema |
| cm |
| firehose |
| hue |
| metastore |
| mysql |
| nav |
| navms |
| oozie |
| retail_db |
| rman |
| sentry |
+-----+
12 rows in set (0.03 sec)

mysql> 

```

3. Now we are using the existing database i.e. retail_db which are already present in mysql.

`use retail_db;`

```
cloudera@quickstart:~$ show databases
+-----+
| Database |
+-----+
| information_schema |
| cm |
| firehose |
| hue |
| metastore |
| mysql |
| nav |
| nvens |
| oozie |
| retail_db |
| rman |
| sentry |
+-----+
12 rows in set (0.03 sec)

mysql> use retail_db;
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_retail_db |
+-----+
| categories |
| customers |
| departments |
| order_items |
| orders |
| products |
+-----+
6 rows in set (0.00 sec)
```

So right now we are under **retail_db** database.

4. Now to see the tables under a specific database so we will be using the same command which is used to display the databases.

`show tables;`

```
cloudera@quickstart:~$ show databases
+-----+
| Database |
+-----+
| rman |
| sentry |
+-----+
12 rows in set (0.03 sec)

mysql> use retail_db;
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_retail_db |
+-----+
| categories |
| customers |
| departments |
| order_items |
| orders |
| products |
+-----+
6 rows in set (0.00 sec)
```

5. Here we are displaying all the records present in customers table using below command.

`select * from customers;`

ID	First Name	Last Name	Address	City	State	Zip Code
98247	Hannah	Brown	8316 Pleasant Bend	Caguas	PR	XXXXXX
09729	Mary	Rios	1221 Cinder Pines	Kaneohe	HI	XXXXXX
96744	Angela	Smith	1525 Jagged Barn Highlands	Caguas	PR	XXXXXX
08725	Benjamin	Garcia	5459 Noble Brook Landing	Levittown	NY	XXXXXX
11756	Mary	Mills	9720 Colonial Parade	Caguas	PR	XXXXXX
69725	Lauren	Horton	5736 Honey Downs	Summerville	SC	XXXXXX
29483						
12435						

12435 rows in set (0.17 sec)

As it is a huge table. It contains total 12435 rows or records.

select * from customers limit 10;

```
cloudera@quickstart:~$ mysql> select * from customers limit 10;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right syntax to use near 'from customers' at line 1
mysql> select * from customers;
+-----+-----+-----+-----+-----+
| customer_id | customer_fname | customer_lname | customer_email | customer_password |
| customer_street | customer_city | customer_state | customer_zipcode |
+-----+-----+-----+-----+-----+
| 1 | Richard | Hernandez | XXXXXXXXX | XXXXXXXXX | |
| 6303 Heather Plaza | Brownsville | TX | 78521 |
| 2 | Mary | Barrett | Littleton | CO | XXXXXXXXX |
| 9526 Noble Embers Ridge | | | | 80126 |
+-----+-----+-----+-----+-----+
```

6. Let see any other table. Here we are displaying department table it has 6 rows in it.

select * from departments;

```
cloudera@quickstart:~$ mysql> select * from departments;
+-----+-----+
| department_id | department_name |
+-----+-----+
| 2 | Fitness |
| 3 | Footwear |
| 4 | Apparel |
| 5 | Golf |
| 6 | Outdoors |
| 7 | Fan Shop |
+-----+-----+
```

These are the different departments i.e. department_name with their respective department_id.

7. Open the new terminal for running command for sqoop.

8. We will require hostname for this sqoop .

hostname -f

```
cloudera@quickstart:~$ hostname -f
quickstart.cloudera
[cloudera@quickstart ~]$
```

So it is giving as **quickstart.cloudera** as the name of the host that we are already connected to.

9. Then if we want to list down all the databases we will use below sqoop command.

sqoop list-databases --connect jdbc:mysql://quickstart:3306 --password cloudera --username root;

So we have studied in this sqoop that we can make use of jdbc or the odbc type driver. So the applications that support the jdbc will be connecting them with the jdbc driver. So here we are using mysql which we are going to connect this with the jdbc . mysql running on **quickstart:3306** then we have to mention password which is **cloudera** and username i.e. **root**.

```
cloudera@quickstart:~$ sqoop list-databases --connect jdbc:mysql://quickstart:3306 --password cloudera --username root;
Warning: /usr/lib/sqoop/.accumulo does not exist. Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/03/25 20:45:11 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
22/03/25 20:45:11 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/03/25 20:45:12 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
information_schema
firehose
hue
metastore
mysql
navs
oozie
retail_db
rman
sentry
[cloudera@quickstart ~]$
```

These are the lists of same databases which are present in mysql and the information also present here. So we are connecting the sqoop with mysql with the help of jdbc.

10. Now we will list out all the tables using below command.

sqoop list-tables --connect jdbc:mysql://quickstart:3306/retail_db --password cloudera --username root;

```
[cloudera@quickstart ~]$ sqoop list-tables --connect jdbc:mysql://quickstart:3306/retail_db --password cloudera --username root;
Warning: /usr/lib/sqoop/.accumulo does not exist. Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/03/25 20:50:38 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.13.0
22/03/25 20:50:38 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/03/25 20:50:31 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
categories
customers
departments
order_items
orders
products
[cloudera@quickstart ~]$
```

11. Now we will start with the import and export tools of the Hadoop. We want to Import table "departments" from retail_db database which are present inside in mysql.

sqoop import --connect jdbc:mysql://quickstart:3306/retail_db --password cloudera --username root --table departments;

```
cloudera@quickstart:~$ sqoop import --connect jdbc:mysql://quickstart:3306/retail_db --password cloudera --username root --table departments
HDFS: Number of large read operations=0
HDFS: Number of write operations=8
Job Counters
Launched map tasks=4
Other local map tasks=0
Total time spent by all maps in occupied slots (ms)=56487424
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=110327
Total vcore-milliseconds taken by all map tasks=110327
Total megabyte-milliseconds taken by all map tasks=56487424
Map-Reduce Framework
Map input records=6
Map output records=6
Input split bytes=481
Spilled Records=0
Failed Shuffles=0
Merged Map outputs=1
CPU time spent (ms)=9887
CPU time spent (ms)=9109
Physical memory (bytes) snapshot=545624064
Virtual memory (bytes) snapshot=2908661696
Total committed heap usage (bytes)=197132288
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=0
22/03/25 20:54:38 INFO mapreduce.ImportJobBase: Transferred 60 bytes in 100.9359 seconds (0.5944 bytes/sec)
22/03/25 20:54:38 INFO mapreduce.ImportJobBase: Retrieved 6 records.
[cloudera@quickstart ~]$
```

12. Now we will see whether all departments table successfully imported from mysql in Hadoop (hdfs) or not using below command.

hadoop fs -ls

```
[cloudera@quickstart ~]$ hadoop fs -ls
Found 3 items
drwx-----  cloudera cloudera  0 2022-03-25 20:54 ,staging
drwxr-xr-x  cloudera cloudera  0 2022-03-25 02:22 Training
drwxr-xr-x  cloudera cloudera  0 2022-03-25 20:54 departments
```

Departments table is now successfully imported in hdfs.

13. Now we will see what inside this department using below command.

hadoop fs -ls departments;

```
[cloudera@quickstart ~]$ hadoop fs -ls departments;
Found 5 items
-rw-r--r--  1 cloudera cloudera  0 2022-03-25 20:54 departments/_SUCCESS
-rw-r--r--  1 cloudera cloudera 21 2022-03-25 20:54 departments/part-m-00000
-rw-r--r--  1 cloudera cloudera 10 2022-03-25 20:54 departments/part-m-00001
-rw-r--r--  1 cloudera cloudera  7 2022-03-25 20:54 departments/part-m-00002
-rw-r--r--  1 cloudera cloudera 22 2022-03-25 20:54 departments/part-m-00003
```

As we can see there are once SUCCESS file and four part-m files which has got the output present inside the departments.

14. Now we will see what there inside this some of the part m files so that can be done with the help of below commands.

hadoop fs -cat /user/cloudera/departments/part-m-00000

hadoop fs -cat /user/cloudera/departments/part-m-00001

hadoop fs -cat /user/cloudera/departments/part-m-00002

hadoop fs -cat /user/cloudera/departments/part-m-00003

```
[cloudera@quickstart ~]$ hadoop fs -ls departments;
Found 5 items
-rw-r--r--  1 cloudera cloudera  0 2022-03-25 20:54 departments/_SUCCESS
-rw-r--r--  1 cloudera cloudera 21 2022-03-25 20:54 departments/part-m-00000
-rw-r--r--  1 cloudera cloudera 16 2022-03-25 20:54 departments/part-m-00001
-rw-r--r--  1 cloudera cloudera  7 2022-03-25 20:54 departments/part-m-00002
-rw-r--r--  1 cloudera cloudera 22 2022-03-25 20:54 departments/part-m-00003
[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/departments/part-m-00000
cat: '/user/cloudera/departments/part-m-00000': No such file or directory
[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/departments/part-m-00000
2.Fitness
3.Footwear
[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/departments/part-m-00001
4.Apparel
[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/departments/part-m-00002
5.Golf
[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/departments/part-m-00003
6.Outdoors
7.Fan Shop
[cloudera@quickstart ~]$
```

15. If want to display output of all part-m files together we will use below command.

hadoop fs -cat /user/cloudera/departments/part*

```
[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/departments/part*
2.Fitness
3.Footwear
4.Apparel
5.Golf
6.Outdoors
7.Fan Shop
[cloudera@quickstart ~]$
```

16. If we want to mention that where should we have our this output in the hdfs so for that we have to mention the target directory.

We will want to import my department table in –target directory as department1.

```
sqoop import --connect jdbc:mysql://quickstart:3306/retail_db --password cloudera --username root --table departments --target-dir /user/cloudera/department1
```

```
cloudera@quickstart-virtualbox:~$ sqoop import --connect jdbc:mysql://quickstart:3306/retail_db --password cloudera --username root --table departments --target-dir /user/cloudera/department1
cloudera@quickstart:~$ Fri Mar 25, 9:28 PM cloudera
File Edit View Search Terminal Help
HDFS: Number of bytes read=481
HDFS: Number of bytes written=60
HDFS: Number of read operations=16
HDFS: Number of large read operations=0
HDFS: Number of write operations=8
Job Counters
Launched map tasks=4
Other local map tasks=4
Total time spent by all maps in occupied slots (ms)=52659712
Total time spent by all reduces in occupied slots (ms)=0
Total time spent by all map tasks (ms)=102051
Total vcore-milliseconds taken by all map tasks=102051
Total megabyte-milliseconds taken by all map tasks=52659712
Map - Reducer Information
Map Input Records=6
Map Output Records=0
Input split bytes=481
Spilled Records=0
Failed Shuffles=0
Map-Reduce Local Shuffles=0
GC time elapsed (ms)=839
CPU time spent (ms)=9508
Physical memory (bytes) snapshot=538230784
Virtual memory (bytes) snapshot=2969560882
Total committed heap usage (bytes)=197132288
File Input Format Counters
Bytes Read=0
File Output Format Counters
Bytes Written=60
22/03/25 21:28:15 INFO mapreduce.ImportJobBase: Transferred 60 bytes in 85.5797 seconds (0.7611 bytes/sec)
22/03/25 21:28:15 INFO mapreduce.ImportJobBase: Retrieved 6 records.
[cloudera@quickstart ~]$
```

As we can see 6 records are retrieved successfully.

17. Now let's check it using below command.

```
hadoop fs -ls
```

```
[cloudera@quickstart ~]$ hadoop fs -ls
Found 4 items
drwx-----  - cloudera cloudera      0 2022-03-25 21:28 .staging
drwxr-xr-x  - cloudera cloudera      0 2022-03-25 02:22 Training
drwxr-xr-x  - cloudera cloudera      0 2022-03-25 21:28 department1
drwxr-xr-x  - cloudera cloudera      0 2022-03-25 20:54 departments
[cloudera@quickstart ~]$
```

So here we have department1 directory.

18. Now we will check what is present inside this department1 directory using below command.

```
hadoop fs -ls /user/cloudera/department1
```

```
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/department1
Found 4 items
drwx-----  - cloudera cloudera      0 2022-03-25 21:28 .staging
drwxr-xr-x  - cloudera cloudera      0 2022-03-25 02:22 Training
drwxr-xr-x  - cloudera cloudera      0 2022-03-25 21:28 department1
drwxr-xr-x  - cloudera cloudera      0 2022-03-25 20:54 departments
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/department1
Found 5 items
-rw-r--r--  1 cloudera cloudera      0 2022-03-25 21:28 /user/cloudera/department1/_SUCCESS
-rw-r--r--  1 cloudera cloudera     21 2022-03-25 21:27 /user/cloudera/department1/part-m-00001
-rw-r--r--  1 cloudera cloudera    10 2022-03-25 21:27 /user/cloudera/department1/part-m-00001
-rw-r--r--  1 cloudera cloudera      7 2022-03-25 21:28 /user/cloudera/department1/part-m-00002
-rw-r--r--  1 cloudera cloudera    22 2022-03-25 21:28 /user/cloudera/department1/part-m-00002
[cloudera@quickstart ~]$
```

So it has different part-m files.

19. Now we will read the content of these part-m files using below command.

```
hadoop fs -cat /user/cloudera/department1/part*
```

```
[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/department1/part*
2,Fitness
3,Footwear
4,Apparel
5,Golf
6,Outdoors
7,Fan Shop
[cloudera@quickstart ~]$
```

So we have got the output.

20. Now we will filter out some or specific rows only from the departments table and have it in hdfs but

before we will apply some conditions on the rows of the departments table and whichever rows will satisfy the condition only those rows are would be stored in the hdfs.

We want to fetch only those departments where department_id is greater than 4.

```
sqoop import --connect jdbc:mysql://quickstart:3306/retail_db --password cloudera --username root --table departments --where "department_id>4" --target-dir /user/cloudera/department2
```

```
HDFS: Number of bytes read=361
HDFS: Number of blocks read=1
HDFS: Number of read operations=12
HDFS: Number of large read operations=0
HDFS: Number of file operations=0
Job Counter
  Job Launched map tasks=2
    Other local map tasks=3
    Total map tasks=5
    Maps in occupied slots (ms)=34762752
    Total time spent by all reduces in occupied slots (ms)=0
    Total time spent by all map tasks (ms)=34762752
    Total memory allocated by all map tasks=0
    Total megabytes-milliseconds taken by all map tasks=0
    Map Framework
      Map input records=3
      Map splits=1
      Input split bytes=361
      Spills=0
      Failed Shuffles=0
      Merged Map inputs=0
      Output file stoppers (ms)=942
      CPU time spent (ms)=610
      Physical memory (bytes) snapshot=412269960
      Virtual memory (bytes) snapshot=218012340
      Total megabytes used by map tasks=147404216
File Input Format Counters
  File Input Format Counters
    Bytes written=0
File Output Format Counters
  Bytes written=25
22/03/25 21:38:29 INFO mapreduce.ImportJobBase: Transferred 25 bytes in 71.7504 seconds (0.4042 bytes/sec)
22/03/25 21:38:29 INFO mapreduce.ImportJobBase: Retrieved 3 records.
[cloudera@quickstart ~]$
```

As we can see 3 records are retrieved successfully.

21. Now we will check it using below command.

```
hadoop fs -ls /user/cloudera/department2
```

```
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/department2
Found 4 items
-rw-r--r-- 1 cloudera cloudera 0 2022-03-25 21:38 /user/cloudera/department2/ SUCCESS
-rw-r--r-- 1 cloudera cloudera 7 2022-03-25 21:38 /user/cloudera/department2/part-m-00000
-rw-r--r-- 1 cloudera cloudera 11 2022-03-25 21:38 /user/cloudera/department2/part-m-00001
-rw-r--r-- 1 cloudera cloudera 11 2022-03-25 21:38 /user/cloudera/department2/part-m-00002
[cloudera@quickstart ~]$
```

22. Now will read the content of these part-m files using cat command.

```
hadoop fs -ls /user/cloudera/department2/part*
```

```
[cloudera@quickstart ~]$ hadoop fs -ls /user/cloudera/department2/part*
-rw-r--r-- 1 cloudera cloudera 7 2022-03-25 21:38 /user/cloudera/department2/part-m-00000
-rw-r--r-- 1 cloudera cloudera 11 2022-03-25 21:38 /user/cloudera/department2/part-m-00001
-rw-r--r-- 1 cloudera cloudera 11 2022-03-25 21:38 /user/cloudera/department2/part-m-00002
[cloudera@quickstart ~]$
```

23. Now we will see the Export command. So what the export tool does is it willexport the data from our hdfs to the RDBMS. So for that we need to have some table in mysql with some records so for that we will now move to mysql.

24. So here we will create the table “dpt” and it will be having two attributes as

“department_id” and “ department_name”.

```
create table dpt(department_id int not null auto_increment, department_name varchar(50) not null, primary key(department_id));
```

```
cloudera@quickstart-vm-3.10.0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~$ Fri Mar 25, 9:48 PM cloudera
File Edit View Search Terminal Help
Total time: 0s
Total memory: 0B
Total memory: 0B
Map-Reduce Frame
  Map Input
  Map Output
  Input split
  Spilled
  Failed
  Merged
  GC time: 0 rows in set (0.00 sec)
  CPU time
  Physical
  Virtual
Total count: 0
File Input Format
  Bytes Read
  File Output Format
  Bytes Written
22/03/25 21:38:29 INFO m
22/03/25 21:38:29 INFO m
[cloudera@quickstart ~]$ mysql> select * from departments;
+-----+-----+
| department_id | department_name |
+-----+-----+
| 2 | Fitness |
| 3 | Clothing |
| 4 | Apparel |
| 5 | Golf |
| 6 | Outdoors |
| 7 | Fan Shop |
+-----+-----+
6 rows in set (0.00 sec)

mysql> Create table dpt(department_id int not null auto increment, department_name varchar(50) not null, primary key(department_id));
Query OK, 0 rows affected (0.06 sec)

[cloudera@quickstart ~]$ hadoop fs -cat /user/cloudera/department2/part*
-rw-r--r-- 1 cloudera cloudera 7 2022-03-25 21:38 /user/cloudera/department2/part-m-00000
-rw-r--r-- 1 cloudera cloudera 11 2022-03-25 21:38 /user/cloudera/department2/part-m-00001
-rw-r--r-- 1 cloudera cloudera 11 2022-03-25 21:38 /user/cloudera/department2/part-m-00002
[cloudera@quickstart ~]$
```

25. Now we want to check what we have inside this dpt table.

select * from dpt;

```
FOUND 4 ITEMS
-rw-r--r-- 1 cloudera mysql  select * from dpt;
Empty set (0.00 sec)
-rw-r--r-- 1 cloudera mysql  |
[cloudera@quickstart ~]$ mysql> |
Empty set (0.00 sec)
7 2022-03-25 21:38 /user/cloudera/department2/part-m-00000
11 2022-03-25 21:38 /user/cloudera/department2/part-m-00001
11 2022-03-25 21:38 /user/cloudera/department2/part-m-00002
[cloudera@quickstart ~]$
```

As we have not inserted any records inside the dpt table so that's why it is showing as Empty set.

26. Now we will exporting the data from the hdfs to dpt table of mysql. Now we will move to the sqoop terminal.

27. So now we are performing export operation using below command.

we are trying to export department2 which are present in cloudera to inside our dpt table which are present inside mysql.

sqoop export --connect jdbc:mysql://quickstart:3306/retail_db --password cloudera --username root --table dpt --export-dir /user/cloudera/department2

```
cloudera@quickstart:~$ 13:00 Oracle VM VirtualBox
File Edit View Search Terminal Help
HDFS: Number of bytes read=0
HDFS: Number of bytes written=0
HDFS: Number of read operations=0
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
Job Counters:
Launched map tasks=3
Data Locality: Local=3
Total time spent by all maps in occupied slots (ms)=32715776
Total time spent by all reducers in occupied slots (ms)=0
Total memory consumed (bytes)=109998
Total vcore-milliseconds taken by all map tasks=63898
Total megabyte-milliseconds taken by all map tasks=32715776
Map Reducer counters:
Map input records=3
Map output records=3
Total map memory used (bytes)=627
Spilled Records=9
Failed Shuffles=0
Map input record bytes=0
GC time elapsed (ms)=558
CPU time spent (s)=0.5946
Physical memory snapshot=412072008
Virtual memory (bytes) snapshot=2109742088
Total committed heap usage (bytes)=14703488
File Input Format Counters:
Bytes Read=0
File Output Format Counters:
22/03/25 21:57:21 INFO mapreduce.ExportJobBase: Transferred 683 bytes in 70.2615 seconds (9.7208 bytes/sec)
22/03/25 21:57:21 INFO mapreduce.ExportJobBase: Exported 3 records.
[cloudera@quickstart ~]$
```

Now we have successfully exported 3 records.

28. Now we will see whether the records are successfully exported in dpt table which are present inside mysql using below command.

select * from dpt;

```
cloudera@quickstart:~$ 13:00 Oracle VM VirtualBox
File Edit View Search Terminal Help
File Applications Places System
cloudera@quickstart:~$ cloudera@quickstart:-
File Edit View Search Terminal Help
HDFS: Number of bytes read=0
HDFS: Number of bytes written=0
HDFS: Number of read operations=0
HDFS: Number of large read operations=0
HDFS: Number of write operations=0
Job Counters:
Launched map tasks=1
Data-local
Total map memory used (bytes)=0
Total vcore-milliseconds taken by all map tasks=0
Total megabyte-milliseconds taken by all map tasks=0
Map Reducer counters:
Map input selected from dpt:
Map input records=3
Map output records=3
Total map memory used (bytes)=0
Map output bytes=0
Spilled Records=0
Failed Shuffles=0
Map input record bytes=0
GC time elapsed (ms)=0
CPU time spent (s)=0.0000
Physical memory snapshot=0
Virtual memory (bytes) snapshot=0
Total committed heap usage (bytes)=0
File Input Format Counters:
Bytes Read=0
File Output Format Counters:
Bytes Written=0
22/03/25 21:57:21 INFO mapreduce.ExportJobBase: Transferred 683 bytes in 70.2615 seconds (9.7208 bytes/sec)
22/03/25 21:57:21 INFO mapreduce.ExportJobBase: Exported 3 records.
[cloudera@quickstart ~]$
```

As we can see these 3 records which are present in department2 table are successfully exported inside the dpt table of mysql.

quit

```
VIRTUAL
Total com.mysql> quit
File Input Format Bye
Bytes Read [cloudera@quickstart ~]$ 
File Output Format Counters
Bytes Written=0
22/03/25 21:57:21 INFO mapreduce.ExportJobBase: Transferred 683 bytes in 70.2615 seconds (9.7208 bytes/sec)
22/03/25 21:57:21 INFO mapreduce.ExportJobBase: Exported 3 records.
[cloudera@quickstart ~]$
```

Aim - Schema design using HBaseWhat is HBase?

HBase is a column-oriented non-relational database management system that runs on top of Hadoop Distributed File System (HDFS). HBase provides a fault-tolerant way of storing sparse data sets, which are common in many big data use cases. It is well suited for real-time data processing or random read/write access to large volumes of data.

Unlike relational database systems, HBase does not support a structured query language like SQL; in fact, HBase isn't a relational data store at all. HBase applications are written in Java™ much like a typical Apache MapReduce application. HBase does support writing applications in Apache Avro, REST and Thrift.

An HBase system is designed to scale linearly. It comprises a set of standard tables with rows and columns, much like a traditional database. Each table must have an element defined as a primary key, and all access attempts to HBase tables must use this primary key.

Avro, as a component, supports a rich set of primitive data types including: numeric, binary data and strings; and a number of complex types including arrays, maps, enumerations and records. A sort order can also be defined for the data.

HBase relies on ZooKeeper for high-performance coordination. ZooKeeper is built into HBase, but if you're running a production cluster, it's suggested that you have a dedicated ZooKeeper cluster that's integrated with your HBase cluster.

HBase works well with Hive, a query engine for batch processing of big data, to enable fault-tolerant big data applications.

An example of HBase

An HBase column represents an attribute of an object; if the table is storing diagnostic logs from servers in your environment, each row might be a log record, and a typical column could be the timestamp of when the log record was written, or the server name where the record originated.

HBase allows for many attributes to be grouped together into column families, such that the elements of a column family are all stored together. This is different from a row-oriented relational database, where all the columns of a given row are stored together. With HBase you must predefine the table schema and specify the column families. However, new columns can be added to families at any time, making the schema flexible and able to adapt to changing application requirements.

Just as HDFS has a NameNode and slave nodes, and MapReduce has JobTracker and TaskTracker slaves, HBase is built on similar concepts. In HBase a master node manages the cluster and region servers store portions of the tables and perform the work on the data. In the same way HDFS has some enterprise concerns due to the availability of the NameNode HBase is also sensitive to the loss of its master node.

HBase Shell

HBase contains a shell using which you can communicate with HBase. HBase uses the Hadoop File System to store its data. It will have a master server and region servers. The data storage will be in the form of regions (tables). These regions will be split up and stored in region servers.

The master server manages these region servers and all these tasks take place on HDFS. Given below are some of the commands supported by HBase Shell.

Before performing the practical first perform 3 steps of the given following

1. sudo /home/cloudera/cloudera-manager --force --express

```
cloudera-quickstart-vm-5.13.0-0-virtualbox (Running) - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~ Fri Mar 25, 1:15 AM cloudera
File Edit View Search Terminal Help
[QuickStart] Waiting for Cloudera Manager API...
[QuickStart] Starting Cloudera Manager agent...
[QuickStart] Configuring deployment...
Submitted jobs: 41
[QuickStart] Deploying client configuration...
Submitted jobs: 42
[QuickStart] Starting Cloudera Management Service...
Submitted jobs: 58
[QuickStart] Enabling Cloudera Manager daemons on boot...

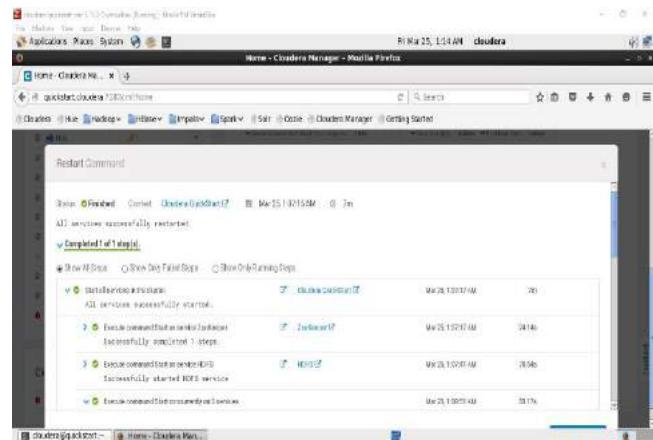
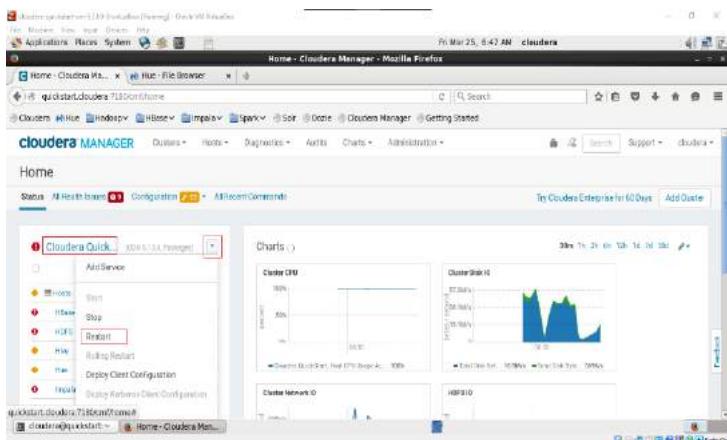
Success! You can now log into Cloudera Manager from the QuickStart VM's browser:
http://quickstart.cloudera:7180

Username: cloudera
Password: cloudera

[cloudera@quickstart ~]$ sudo -u hdfs hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

22/03/25 01:04:22 WARN ipc.Client: Failed to connect to server: quickstart.cloudera/10.0.2.15:8820; try once and fail.
java.net.ConnectException: Connection refused
    at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
    at sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:739)
    at org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout.java:206)
    at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:538)
    at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:494)
    at org.apache.hadoop.ipc.Client$Connection.setupConnection(Client.java:648)
    at org.apache.hadoop.ipc.Client$Connection.access$3000(Client.java:744)
    at org.apache.hadoop.ipc.Client$Connection.access$3000(Client.java:396)
```

2. Start all services



```
cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~ Fri Mar 25, 1:16 AM cloudera
File Edit View Search Terminal Help
at org.apache.hadoop.ipc.Client$Connection.setupConnection(Client.java:648)
at org.apache.hadoop.ipc.Client$Connection.access$3000(Client.java:396)
at org.apache.hadoop.ipc.Client.getConnection(Client.java:1557)
at org.apache.hadoop.ipc.Client.call(Client.java:1488)
at org.apache.hadoop.ipc.Client.call(Client.java:1446)
at org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngine.java:230)
at com.sun.proxy.$Proxy16.setSafeMode(Unknown Source)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:608)
at org.apache.hadoop.ipc.RetryInvocationHandler.invoke(RetryInvocationHandler.java:268)
at org.apache.hadoop.ipc.RetryInvocationHandler.invoke(RetryInvocationHandler.java:104)
at com.sun.proxy.$Proxy17.setSafeMode(Unknown Source)
at org.apache.hadoop.hdfs.DFSClient.setSafeMode(DFSClient.java:2648)
at org.apache.hadoop.hdfs.DistributedFileSystem.setSafeMode(DistributedFileSystem.java:1182)
at org.apache.hadoop.hdfs.DistributedFileSystem.setSafeMode(DistributedFileSystem.java:1166)
at org.apache.hadoop.hdfs.tools.DFSAdmin.setSafeMode(DFSAdmin.java:570)
at org.apache.hadoop.hdfs.tools.DFSAdmin.run(DFSAdmin.java:1856)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:84)
at org.apache.hadoop.hdfs.tools.DFSAdmin.main(DFSAdmin.java:2832)
safeMode: Call From quickstart.cloudera[10.6.2.15] to quickstart.cloudera:8020 failed on connection exception: java.net.ConnectException: Connection refused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
[cloudera@quickstart ~]$ sudo -u hdfs hadoop dfsadmin -safeMode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Safe mode is OFF
[cloudera@quickstart ~]$
```

3. sudo -u hdfs hadoop dfsadmin -safemode leave

We can start the HBase interactive shell using “hbase shell” command as shown below.

hbase shell

```

cloudera@quickstart-vn-5:10.0-VirtualBox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~ Fri Mar 25, 1:21 AM cloudera

File Edit View Search Terminal Help
at org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngine.java:239)
at com.sun.proxy.$Proxy16.setSafeMode(Unknown Source)
at org.apache.hadoop.hdfs.protocolPB.ClientNameNodeProtocolTranslatorPB.setSafeMode(ClientNameNodeProtocolTranslatorPB.java:681)
at sun.reflect.NativeMethodAccessorImpl.invoke(Native Method)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
at java.lang.reflect.Method.invoke(Method.java:606)
at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:260)
at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:104)
at com.sun.proxy.$Proxy17.setSafeMode(Unknown Source)
at org.apache.hadoop.hdfs.DistributedFileSystem.setSafeMode(DistributedFileSystem.java:2048)
at org.apache.hadoop.hdfs.DistributedFileSystem.setSafeMode(DistributedFileSystem.java:1182)
at org.apache.hadoop.hdfs.tools.DFSAdmin.setSafeMode(DFSAdmin.java:576)
at org.apache.hadoop.hdfs.tools.DFSAdmin.run(DFSAdmin.java:1856)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:84)
at org.apache.hadoop.hdfs.DFSAdmin.main(DFSAdmin.java:2032)
safemode: Call From quickstart.cloudera/10.0.2.15 to quickstart.cloudera:8020 failed on connection exception: java.net.ConnectException: Connection refused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
[cloudera@quickstart ~]$ sudo -u hdfs hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Safe mode is OFF
[cloudera@quickstart ~]$ hbase shell
22/03/25 01:19:20 INFO Configuration.deprecation: hadoop.native.lib is deprecated. Instead, use io.native.lib.available
HBase Shell; enter 'help|RETURN' for list of supported commands.
Type "exit|UNKNOWN" to leave the HBase Shell.
Version 1.2.0-cdh5.13.0, runknown, Wed Oct 4 11:16:18 PDT 2017
hbase(main):001:0> 
```

Check the shell functioning before proceeding further. Use the list command for this purpose. List is a command used to get the list of all the tables in HBase. It lists all the tables in HBase.

List

```

cloudera@quickstart-vn-5:10.0-VirtualBox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~ Fri Mar 25, 1:22 AM cloudera

File Edit View Search Terminal Help
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:606)
at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:260)
at org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocationHandler.java:104)
at com.sun.proxy.$Proxy17.setSafeMode(Unknown Source)
at org.apache.hadoop.hdfs.DistributedFileSystem.setSafeMode(DistributedFileSystem.java:2048)
at org.apache.hadoop.hdfs.DistributedFileSystem.setSafeMode(DistributedFileSystem.java:1182)
at org.apache.hadoop.hdfs.tools.DFSAdmin.setSafeMode(DFSAdmin.java:576)
at org.apache.hadoop.hdfs.tools.DFSAdmin.run(DFSAdmin.java:1856)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:84)
at org.apache.hadoop.hdfs.tools.DFSAdmin.main(DFSAdmin.java:2032)
safemode: Call From quickstart.cloudera/10.0.2.15 to quickstart.cloudera:8020 failed on connection exception: java.net.ConnectException: Connection refused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
[cloudera@quickstart ~]$ sudo -u hdfs hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Safe mode is OFF
[cloudera@quickstart ~]$ hbase shell
22/03/25 01:19:20 INFO Configuration.deprecation: hadoop.native.lib is deprecated. Instead, use io.native.lib.available
HBase Shell; enter 'help|RETURN' for list of supported commands.
Type "exit|UNKNOWN" to leave the HBase Shell.
Version 1.2.0-cdh5.13.0, runknown, Wed Oct 4 11:16:18 PDT 2017
hbase(main):001:0> list
TABLE
0 row(s) in 0.0209 seconds
=> []
hbase(main):002:0> 
```

Restart HBase services if this is not running on terminal

\$ sudo su – This command is to become super user

\$ service hbase-master restart – This command is to restart hbase-master services

\$ service hbase-regionserver restart – This command is to restart hbase-regionserver services

Once these commands run successfully then Open the browser and refresh the page and see all the HBase servers will be restarted.

Schema Design using HBase

The HBase schema design is very different compared to the relation database schema design. Below are some of general concept that should be followed while designing schema in Hbase:

Row key: Each table in HBase table is indexed on row key. Data is sorted lexicographically by this row key. There are no secondary indices available on HBase table.

Automaticity: Avoid designing table that requires atomacity across all rows. All operations on HBase rows are atomic at row level.

Even distribution: Read and write should uniformly distributed across all nodes available in cluster. Design row key in such a way that, related entities should be stored in adjacent rows to increase read efficacy.

HBase Schema Row key, Column family, Column qualifier, individual and Row value SizeLimit

1. Creating a Table using HBase Shell

We can create a table using the create command, here you must specify the table name and the Column Family name. The syntax to create a table in HBase shell is shown below.

create '<table name>','<column family>'

create 'customer','address','order'

customer -> table ; address & order -> column_name

```

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System cloudera@quickstart:~ Fri Mar 25, 1:24 AM cloudera
File Edit View Search Terminal Help
at com.sun.proxy.$Proxy17.setSafeMode(Unknown Source)
at org.apache.hadoop.hdfs.DFSClient.setSafeMode(DFSClient.java:2648)
at org.apache.hadoop.hdfs.DistributedFileSystem.setSafeMode(DistributedFileSystem.java:1182)
at org.apache.hadoop.hdfs.DistributedFileSystem.setSafeMode(DistributedFileSystem.java:1166)
at org.apache.hadoop.hdfs.DFSAdmin.setSafeMode(DFSAdmin.java:576)
at org.apache.hadoop.hdfs.tools.DFSAdmin.run(DFSAdmin.java:1856)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:84)
at org.apache.hadoop.hdfs.tools.DFSAdmin.main(DFSAdmin.java:2032)
safemode: Call From quickstart.cloudera/10.8.2.15 to quickstart.cloudera:8020 Tailed on connection exception: java.net.ConnectException: Connection refused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
[cloudera@quickstart ~]$ sudo -u hdfs hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Safe mode is OFF
[cloudera@quickstart ~]$ hbase shell
22/03/25 01:19:20 INFO Configuration.deprecation: hadoop.native.lib is deprecated. Instead, use io.native.lib.available
HBase Shell; enter 'help|RETURN' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.0-cdh5.13.0, rUNKNOWN, Wed Oct 4 11:16:18 PDT 2017

hbase(main):081:0> list
TABLE
0 row(s) in 0.8200 seconds

=> []
hbase(main):082:0> create 'customer','address','order'
0 row(s) in 2.6180 seconds

=> Hbase::Table - customer
hbase(main):083:0> 

cloudera@quickstart:~$ Home - Cloudera Man...

```

list

```

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System cloudera@quickstart:~ Fri Mar 25, 1:27 AM cloudera
File Edit View Search Terminal Help
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:84)
at org.apache.hadoop.hdfs.tools.DFSAdmin.main(DFSAdmin.java:2032)
safemode: Call From quickstart.cloudera/10.8.2.15 to quickstart.cloudera:8020 failed on connection exception: java.net.ConnectException: Connection refused; For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
[cloudera@quickstart ~]$ sudo -u hdfs hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

Safe mode is OFF
[cloudera@quickstart ~]$ hbase shell
22/03/25 01:19:20 INFO Configuration.deprecation: hadoop.native.lib is deprecated. Instead, use io.native.lib.available
HBase Shell; enter 'help|RETURN' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.0-cdh5.13.0, rUNKNOWN, Wed Oct 4 11:16:18 PDT 2017

hbase(main):001:0> list
TABLE
0 row(s) in 0.8280 seconds

=> []
hbase(main):002:0> create 'customer','address','order'
0 row(s) in 2.6188 seconds

=> Hbase::Table - customer
hbase(main):003:0> list
TABLE
customer
1 row(s) in 0.0450 seconds

=> ['customer']
hbase(main):004:0> 

cloudera@quickstart:~$ Home - Cloudera Man...

```

2. Put: Inserts a new record into the table with row identified by 'row.'

This command is used for following things

- It will put a cell 'value' at defined or specified table or row or column.
- It will optionally coordinate time stamp.

Syntax: put <tablename>,<rowname>,<columnvalue>,<value>

Example – with the help of put commands we have inserted new records in “customer” table for address and order family. Here name “Nick” is Row key

```
put 'customer','Nick','address:city','Mumbai'
put 'customer','Nick','address:state','Maharashtra'
put 'customer','Nick','address:street','Street1'
put 'customer','Nick','order:number','ORD-15'
put 'customer','Nick','order:amount','50'
put 'customer','Nick','address:state','Maharashtra'
```

```
cloudera-quickstart-vm-5.15.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~$ Fri Mar 25, 1:30 AM cloudera
File Edit View Search Terminal Help
safe mode: Call From quickstart.cloudera:19.8.2.15 to quickstart.cloudera:8020 failed on connection exception: java.net.ConnectException: Connection refused. For more details see: http://wiki.apache.org/hadoop/ConnectionRefused
[cloudera@quickstart ~]$ sudo -u hdfs hadoop dfsadmin -safemode leave
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead, use the hdfs command for it.
safe mode is OFF
[cloudera@quickstart ~]$ hbase shell
22/03/25 01:19:20 INFO Configuration.deprecation: hadoop.native.lib is deprecated. Instead, use io.native.lib.available
HBase Shell: enter 'help->RETURN' for list of supported commands.
Type exit() or quit() to leave the HBase Shell.
version 1.2.8-cdh5.13.0, runknown, Wed Oct 4 11:16:18 PDT 2017
hbase(main):001:0> list
TABLE
0 row(s) in 0.0200 seconds
=> []
hbase(main):002:0> create 'customer', 'address', 'order'
0 row(s) in 2.6166 seconds
=> Hbase::Table - customer
hbase(main):003:0> list
TABLE
customer
1 row(s) in 0.0456 seconds
=> ['customer']
hbase(main):004:0> put 'customer', 'Nick', 'address:city', 'Mumbai'
0 row(s) in 0.4296 seconds
hbase(main):005:0> list
[cloudera@quickstart ~] Home - Cloudera Manager - Mozilla Firefox
[cloudera@quickstart ~] Home - Cloudera Manager...
```

Adding one more record of customer name “Justin”. Here name “Justin” is Row key

```
put 'customer','Justin','address:city','Pune'
put 'customer','Justin','address:state','Maharashtra'
put 'customer','Justin','order:number','ORD-16'
put 'customer','Justin','order:amount','60'
```

```
cloudera-quickstart-vm-5.15.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~$ Fri Mar 25, 1:37 AM cloudera
File Edit View Search Terminal Help
0 row(s) in 0.0260 seconds
hbase(main):006:0> put 'customer', 'Nick', 'order:amount', '50'
0 row(s) in 0.0128 seconds
hbase(main):007:0> put 'customer', 'Nick', 'order:number', 'ORD-15'
0 row(s) in 0.0119 seconds
hbase(main):016:0> put 'customer', 'Justin', 'address:city', 'Pune'
0 row(s) in 0.0300 seconds
hbase(main):011:0> put 'customer', 'Justin', 'address:state', 'Maharashtra'
0 row(s) in 0.0088 seconds
hbase(main):012:0> put 'customer', 'Justin', 'address:street', 'street2'
0 row(s) in 0.0099 seconds
hbase(main):013:0> put 'customer', 'Justin', 'order:number', 'ORD-16'
0 row(s) in 0.0100 seconds
hbase(main):014:0> put 'customer', 'Justin', 'order:amount', '60'
0 row(s) in 0.0130 seconds
hbase(main):015:0> list
[cloudera@quickstart ~] Home - Cloudera Manager...
```

3. Get: Returns the records matching the row identifier provided in the table

By using this command, you will get a row or cell contents present in the table. In addition to that you can also add additional parameters to it like TIMESTAMP, TIMERANGE, VERSIONS, FILTERS, etc. to get a particular row or cell content.

Syntax: get <tablename>,<rowname>,{< Additional parameters>}

a) **get 'customer', 'Nick'**

```

cloudera@quickstart:~$ hbase(main):011:0> put 'customer','Justin','address:state','Maharashtra'
0 row(s) in 0.0080 seconds

hbase(main):012:0> put 'customer','Justin','address:street','street2'
0 row(s) in 0.0090 seconds

hbase(main):013:0> put 'customer','Justin','order:number','ORD-16'
0 row(s) in 0.0100 seconds

hbase(main):014:0> put 'customer','Justin','order:amount','68'
0 row(s) in 0.0130 seconds

hbase(main):015:0> get 'customer', 'Nick'
COLUMN          CELL
address:city    timestamp=1648197017890, value=Mumbai
address:number   timestamp=1648197149436, value=ORD-15
address:state    timestamp=1648197068315, value=Maharashtra
address:street   timestamp=1648197090129, value=street1
order:amount     timestamp=1648197193433, value=50
order:number    timestamp=1648197213674, value=ORD-15
6 row(s) in 0.0490 seconds

hbase(main):016:0>

```

b) Additional parameters to get only address details

get 'customer', 'Nick','address'

```

cloudera@quickstart:~$ hbase(main):016:0> get 'customer', 'Justin'
COLUMN          CELL
address:state   timestamp=1648197068315, value=Maharashtra
address:street  timestamp=1648197275891, value=street1
order:amount    timestamp=1648197193433, value=50
order:number   timestamp=1648197213674, value=ORD-15
0 row(s) in 0.0490 seconds

hbase(main):017:0> get 'customer', 'Nick'
COLUMN          CELL
address:city    timestamp=1648197017890, value=Pune
address:state   timestamp=1648197275891, value=Maharashtra
address:street  timestamp=1648197293721, value=street2
order:amount    timestamp=1648197354540, value=60
order:number   timestamp=1648197336814, value=ORD-16
5 row(s) in 0.0260 seconds

hbase(main):017:0> get 'customer', 'Nick', 'address'
COLUMN          CELL
address:city    timestamp=1648197017890, value=Mumbai
address:number   timestamp=1648197149436, value=ORD-15
address:state    timestamp=1648197068315, value=Maharashtra
address:street   timestamp=1648197090129, value=street1
4 row(s) in 0.0220 seconds

hbase(main):018:0>

```

c) Additional parameters to get only city details

get 'customer', 'Nick','address:city'

```

cloudera@quickstart:~$ hbase(main):017:0> get 'customer', 'Justin'
COLUMN          CELL
order:number   timestamp=1648197336814, value=ORD-16
5 row(s) in 0.0260 seconds

hbase(main):017:0> get 'customer', 'Nick', 'address'
COLUMN          CELL
address:city    timestamp=1648197017890, value=Mumbai
address:number   timestamp=1648197149436, value=ORD-15
address:state    timestamp=1648197068315, value=Maharashtra
address:street   timestamp=1648197090129, value=street1
4 row(s) in 0.0220 seconds

hbase(main):018:0> get 'customer', 'Justin', 'address'
COLUMN          CELL
address:city    timestamp=1648197260721, value=Pune
address:state   timestamp=1648197275891, value=Maharashtra
address:street  timestamp=1648197293721, value=street2
3 row(s) in 0.0550 seconds

hbase(main):019:0> get 'customer', 'Nick', 'address:city'
COLUMN          CELL
address:city    timestamp=1648197017890, value=Mumbai
1 row(s) in 0.0480 seconds

hbase(main):020:0>

```

4. Scan: The scan command is used to view the data in HTable. Using the scan

command, you can get the table data.

- This command scans entire table and displays the table contents.
- We can pass several optional specifications to this scan command to get more information about the tables present in the system.
- Scanner specifications may include one or more of the following attributes.
 - These are TIMERANGE, FILTER, TIMESTAMP, LIMIT, MAXLENGTH, COLUMNS, CACHE, STARTROW and STOPROW.

Its syntax is as follows:

Syntax: **scan <'tablename'>, {Optional parameters}**

scan 'customer'

When we execute above commands in HBase then we will be getting all the table “customer” contents along with additional parameters like timestamp as show in below screenshot.

```

cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Fri Mar 25, 1:42 AM cloudera@quickstart:~
File Edit View Search Terminal Help
address:city timestamp=1648197017890, value=Mumbai
1 row(s) in 0.0480 seconds
hbase(main):028:~> scan 'customer'
ROW
  COLUMN=<CELL>
  Justin    column=address:city, timestamp=1648197260721, value=Pune
  Justin    column=address:state, timestamp=1648197275891, value=Maharashtra
  Justin    column=address:street, timestamp=1648197293721, value=street
  Justin    column=order:amount, timestamp=1648197354540, value=60
  Justin    column=order:number, timestamp=1648197336014, value=ORD-16
  Nick      column=address:city, timestamp=1648197017890, value=Mumbai
  Nick      column=address:number, timestamp=1648197149436, value=ORD-15
  Nick      column=address:state, timestamp=1648197068315, value=Maharashtra
  Nick      column=address:street, timestamp=1648197098129, value=street
  Nick      column=order:amount, timestamp=1648197193433, value=50
  Nick      column=order:number, timestamp=1648197213674, value=ORD-15
2 row(s) in 0.1280 seconds
hbase(main):021:~>

```

5. Delete -Using the delete command, you can delete a specific cell in a table.

- This command will delete cell value at defined table of row or column.
- Delete must and should match the deleted cells coordinates exactly.
- When scanning, delete cell suppresses older versions of values. The syntax of delete command is as follows:

Syntax:**delete <'tablename'>,<'row name'>,<'column name'>**

delete 'customer','Nick','address:street'

The above command below delete street from address family for row key “Nick” from “customer” table.

Use scan command to see if street is deleted for customer “Nick”. As we can see “street” information is deleted from customer “Nick” from below screenshot.

```

hbase(main):021:0> delete 'customer','Nick','address:street';
8 row(s) in 0.0770 seconds

hbase(main):022:0> scan 'customer'
ROW
  COLUMN+CELL
  Justin    column=order:amount, timestamp=1648197354540, value=68
  Justin    column=order:number, timestamp=1648197336814, value=ORD-16
  Nick     column=address:city, timestamp=164819717890, value=Mumbai
  Nick     column=address:number, timestamp=1648197149436, value=ORD-15
  Nick     column=address:state, timestamp=1648197068315, value=Maharashtra
  Nick     column=address:street, timestamp=1648197098129, value=street2
  etl      column=order:amount, timestamp=1648197193433, value=50
  Nick     column=order:number, timestamp=1648197213674, value=ORD-15
2 row(s) in 0.1280 seconds

hbase(main):022:0> scan 'customer'
ROW
  COLUMN+CELL
  Justin    column=address:city, timestamp=1648197268721, value=Pune
  Justin    column=address:state, timestamp=1648197275891, value=Maharashtra
  Justin    column=address:street, timestamp=1648197293721, value=street2
  Justin    column=order:amount, timestamp=1648197193433, value=50
  Justin    column=order:number, timestamp=1648197213674, value=ORD-15
  Nick     column=address:city, timestamp=1648197017890, value=Mumbai
  Nick     column=address:number, timestamp=1648197149436, value=ORD-15
  Nick     column=address:state, timestamp=1648197068315, value=Maharashtra
  Nick     column=order:amount, timestamp=1648197193433, value=50
  Nick     column=order:number, timestamp=1648197213674, value=ORD-15
2 row(s) in 0.0598 seconds

hbase(main):023:0>

```

6. Alter - This command alters the column family schema. To understand what exactly it does, we have explained it here with an example.

Alter commands are useful for below cases -

- Altering single, multiple column family names
- Deleting column family names from table
- Several other operations using scope attributes with table

Syntax: **alter <tablename>, NAME=><column family name>, VERSIONS=>5**

We can delete specific column family by using alter commands

alter 'customer','delete' => 'address'

After deleting "address" family from "customer" table. Let's again check customer table using "scan" commands as follow

scan 'customer'

As you can see from below screenshot we only now have order:amount and order:number.

```

hbase(main):023:0> alter 'customer','delete' => 'address';
Updating all regions with the new schema...
8/1 regions updated.
1/1 regions updated.
Done.
8 row(s) in 3.4648 seconds

hbase(main):024:0> scan 'customer'
ROW
  COLUMN+CELL
  Justin    column=order:amount, timestamp=1648197354540, value=68
  Justin    column=order:number, timestamp=1648197336814, value=ORD-16
  Nick     column=order:amount, timestamp=1648197193433, value=50
  Nick     column=order:number, timestamp=1648197213674, value=ORD-15
2 row(s) in 8.0340 seconds

hbase(main):025:0>

```

7. Describe - This command describes the named table.

- It will give more information about column families present in the mentioned table
- In our case, it gives the description about table "customer."
- It will give information about table name with column families, associated filters, versions and some more details.

Syntax:**describe <table name>**

desc 'customer'

```
hbase(main):025:0> desc 'customer'
Table customer is ENABLED
customer
COLUMN FAMILIES DESCRIPTION
{NAME => 'order', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICATION_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN VERSIONS => '0', TTL => 'FO
REVER', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}
1 row(s) in 0.0040 seconds

hbase(main):026:0> ■
```

8. Versions –

A {row, column, version} tuple exactly specifies a cell in HBase. In the Apache HBase you can have many cells where row and columns are same but differs only in version values. A version is a timestamp values is written alongside each value. By default, the timestamp values represent the time on the RegionServer when the data was written, but you can change the default HBase setting and specify a different timestamp value when you put data into the cell.

In HBase, rows and column keys are expressed as bytes, the version is specified using a longinteger. The HBase version dimension is stored in decreasing order, so that when reading from a store file, the most recent values are found first.

create 'customer1',{NAME => 'address', VERSIONS => 3}

With the help of above commands we are creating 3 versions

```
cloudera-quickstart-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Terminal Fri Mar 25, 1:51 AM cloudera
File Edit View Search Terminal Help
Nick column=order:amount, timestamp=1648197193433, value=50
Nick column=order:number, timestamp=1648197213674, value=ORD-15
2 row(s) in 0.0590 seconds

hbase(main):023:0> alter 'customer1','delete' => 'address'
Updating all regions with the new schema...
8/3 regions updated.
1/3 regions updated.
Done.
8 row(s) in 3.4648 seconds

hbase(main):024:0> scan 'customer'
Row column=order:amount, timestamp=1648197354540, value=60
column=order:number, timestamp=164819736014, value=ORD-16
Nick column=order:amount, timestamp=1648197193433, value=50
Nick column=order:number, timestamp=1648197213674, value=ORD-15
2 row(s) in 0.0340 seconds

hbase(main):025:0> desc 'customer'
Table customer is ENABLED
customer
COLUMN FAMILIES DESCRIPTION
{NAME => 'order', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICATION_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN VERSIONS => '0', TTL => 'FO
REVER', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}
1 row(s) in 0.0040 seconds

hbase(main):026:0> create 'customer1',{NAME => 'address', VERSIONS => 3}
8 row(s) in 1.3028 seconds

=> Hbase::Table - customer1
hbase(main):027:0> ■
```

Verifying if “customer1” is created after executing above commands with the help of list commands as shown in screenshot below.

```
cloudera-quickstart-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Terminal Fri Mar 25, 1:52 AM cloudera
File Edit View Search Terminal Help
1/1 regions updated.
Done.
8 row(s) in 3.4640 seconds

hbase(main):024:0> scan 'customer'
Row column=order:amount, timestamp=1648197354540, value=60
column=order:number, timestamp=164819736014, value=ORD-16
Nick column=order:amount, timestamp=1648197193433, value=50
Nick column=order:number, timestamp=1648197213674, value=ORD-15
2 row(s) in 0.0340 seconds

hbase(main):025:0> desc 'customer'
Table customer is ENABLED
customer
COLUMN FAMILIES DESCRIPTION
{NAME => 'order', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICATION_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN VERSIONS => '0', TTL => 'FO
REVER', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}
1 row(s) in 0.0040 seconds

hbase(main):026:0> create 'customer1',{NAME => 'address', VERSIONS => 3}
8 row(s) in 1.3028 seconds

=> Hbase::Table - customer1
hbase(main):027:0> list
TABLE
customer
customer1
2 row(s) in 0.0060 seconds

=> ['customer', 'customer1']
hbase(main):028:0> ■
```

9. Count - You can count the number of rows of a table using the count command. Its syntax is as follows:

count 'customer'

```
hbase(main):027:0> list
TABLE
customer
customer1
2 row(s) in 0.0000 seconds
=> ["customer", "customer1"]
hbase(main):028:0> count 'customer'
2 row(s) in 0.2910 seconds
=> 2
hbase(main):029:0> ■
```

10. Alter -Update the version number for already existing columns family

alter 'customer', NAME => 'address', VERSIONS => 5

```
hbase(main):029:0> alter 'customer', NAME => 'address', VERSIONS => 5
Updating all regions with the new schema...
1/1 regions updated.
Done.
0 row(s) in 2.3090 seconds
hbase(main):030:0> ■
```

Again using describe command to see version is updated

desc 'customer'

```
hbase(main):030:0> desc 'customer'
Table customer is ENABLED
customer
COLUMN FAMILIES DESCRIPTION
{NAME => 'address', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICATION_SCOPE => '0', COMPRESSION => 'NONE', VERSIONS => '5', TTL => 'FOREVER', MIN_VERSION => '0', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}
{NAME => 'order', DATA_BLOCK_ENCODING => 'NONE', BLOOMFILTER => 'ROW', REPLICATION_SCOPE => '0', VERSIONS => '1', COMPRESSION => 'NONE', MIN_VERSIONS => '0', TTL => 'FOREVER', KEEP_DELETED_CELLS => 'FALSE', BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}
2 row(s) in 0.0710 seconds
hbase(main):031:0> ■
```

scan 'customer'

```
hbase(main):031:0> scan 'customer'
ROW                                         COLUMN+CELL
Justin                                         column=address:city, timestamp=1648197260721, value=Pune
Justin                                         column=address:state, timestamp=1648197275891, value=Maharashtra
Justin                                         column=address:street, timestamp=1648197293721, value=street2
Justin                                         column=order:amount, timestamp=1648197354540, value=60
Justin                                         column=order:number, timestamp=1648197336614, value=ORD-16
Nick                                           column=address:city, timestamp=1648197817899, value=Mumbai
Nick                                           column=address:number, timestamp=1648197149436, value=ORD-15
Nick                                           column=address:state, timestamp=1648197068315, value=Maharashtra
Nick                                           column=order:amount, timestamp=1648197193433, value=50
Nick                                           column=order:number, timestamp=1648197213674, value=ORD-15
2 row(s) in 0.0918 seconds
hbase(main):032:0> ■
```

put 'customer', 'Nick' , 'address:city', 'Pune'

put 'customer', 'Nick' , 'address:city', 'Bangalore'

put 'customer', 'Nick' , 'address:city', 'Delhi'

```
hbase(main):032:0> put 'customer','Nick','address:city','Pune'
0 row(s) in 0.0138 seconds
hbase(main):033:0> put 'customer','Nick','address:city','Bangalore'
0 row(s) in 0.0140 seconds
hbase(main):034:0> put 'customer','Nick','address:city','Delhi'
0 row(s) in 0.0190 seconds
hbase(main):035:0> scan 'customer'
ROW                                         COLUMN+CELL
Justin                                         column=address:city, timestamp=1648197260721, value=Pune
Justin                                         column=address:state, timestamp=1648197275891, value=Maharashtra
Justin                                         column=address:street, timestamp=1648197293721, value=street2
Justin                                         column=order:amount, timestamp=1648197354540, value=60
Justin                                         column=order:number, timestamp=1648197336614, value=ORD-16
Nick                                           column=address:city, timestamp=1648197149436, value=Delhi
Nick                                           column=address:number, timestamp=1648197149436, value=ORD-15
Nick                                           column=address:state, timestamp=1648197060315, value=Maharashtra
Nick                                           column=order:amount, timestamp=1648197193433, value=50
Nick                                           column=order:number, timestamp=1648197213674, value=ORD-15
2 row(s) in 0.0289 seconds
hbase(main):036:0> ■
```

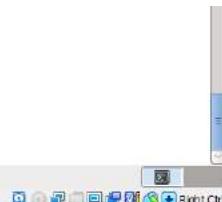
Below are example of adding version of address:city

Currently below is the data in 'customer'. We will be adding more information in it one by one.

scan 'customer', {COLUMN => 'address:city', VERSIONS => 2}

```
hbase(main):036:0> scan 'customer', {COLUMN => 'address:city', VERSIONS => 2}
ROW                                     COLUMN+CELL
Justin                                column=address:city, timestamp=1648197260721, value=Pune
Nick                                  column=address:city, timestamp=1648198762680, value=Delhi
Nick                                  column=address:city, timestamp=1648198755614, value=Bangalore
2 row(s) in 0.0260 seconds
```

```
hbase(main):037:0> 
```



scan 'customer', {COLUMN=> 'address:city', VERSIONS => 3}

scan 'customer', {COLUMN=> 'address:city', VERSIONS => 4}

After executing above commands its giving us 3 and 4 version of address:city records as shown in below screenshot.

scan 'customer', {VERSIONS => 5}

```
hbase(main):042:0> scan 'customer', {VERSIONS => 5}
ROW                                     COLUMN+CELL
Justin                                column=address:city, timestamp=1648197260721, value=Pune
Justin                                column=address:state, timestamp=1648197275891, value=Maharashtra
Justin                                column=address:street, timestamp=1648197293721, value=street2
Justin                                column=order:amount, timestamp=1648197354540, value=50
Justin                                column=order:number, timestamp=1648197360014, value=ORD-16
Nick                                  column=address:city, timestamp=1648198762680, value=Delhi
Nick                                  column=address:city, timestamp=1648198755614, value=Bangalore
Nick                                  column=address:city, timestamp=1648198734145, value=Pune
Nick                                  column=address:city, timestamp=1648198717890, value=Mumbai
Nick                                  column=address:number, timestamp=1648197149436, value=ORD-15
Nick                                  column=address:state, timestamp=1648197868315, value=Maharashtra
Nick                                  column=order:amount, timestamp=1648197183433, value=50
Nick                                  column=order:number, timestamp=1648197213674, value=ORD-15
2 row(s) in 0.0760 seconds
```

```
hbase(main):043:0> 
```

11. Disable -This command will start disabling the named table

If table needs to be deleted or dropped, it has to disable first

Syntax: **disable <table_name>**

disable 'customer'

```
=> ["customer", "customer1"]
hbase(main):047:0> disable 'customer'
8 row(s) in 0.0720 seconds
```



12. Drop– It drops a table from HBase. Drop means complete deletion of table. For this first disable the table then drop it.

- a. To delete the table present in HBase, first we have to disable it
- b. To drop the table present in HBase, first we have to disable it
- c. So either table to drop or delete first the table should be disable using disable command

Syntax:**drop <table_name>**

drop 'customer'

```
hbase(main):047:0> disable 'customer'
8 row(s) in 0.0728 seconds

hbase(main):048:0> drop 'customer'
8 row(s) in 1.3838 seconds

hbase(main):049:0> list
TABLE
customer1
1 row(s) in 0.0228 seconds

=> ["customer1"]
hbase(main):050:0>
```

