BIGDATA

SR NO	NAME OF THE EXPERIMENT	DATE	FACULTY SIGN	CO
1	HDFS: List of Commands (mkdir, touchz, copy from local/put, copy to local/get, move from local, cp, rmr, du, dus, stat)			CO1
2	Map Reduce: 1. Write a program in Map Reduce for Word Count operation. 2. Write a program in Map Reduce for Union operation. 3. Write a program in Map Reduce for Intersection operation. 4. Write a program in Map Reduce for Matrix Multiplication.			CO2
3	MongoDB: 1. Installation 2. Sample Database Creation 3. Query the Sample Database using MongoDB querying mands a. Create Collection b. Insert Document c. Query Document d. Delete Document e. Indexing			CO3

<table>
<tr><td>4</td><td>Hive:
1. Hive Data Types
2. Create Database & Table in Hive
3. Hive Partitioning
4. Hive Built-In Operators
5. Hive Built-In Functions
6. Hive Views
7. HiveQL: Select Where, Select OrderBy, Select GroupBy,
Select Joins</td><td></td><td></td><td>CO4</td></tr>
<tr><td>5</td><td>Pig:
1. Pig Latin Basic
2. Pig Data Types,
3. Download the data
4. Create your Script
5. Save and Execute the Script
6. Pig Operations: Diagnostic Operators, Grouping and Joining, Combining & Splitting, Filtering, Sorting</td><td></td><td></td><td>CO4</td></tr>
<tr><td>6</td><td>Spark:<table><tr><td>Apache Spark
Commands in Scala</td><td>Pair RDD
(Key-Value RDD)
Operations</td></tr><tr><td></td><td></td></tr><tr><td>Start Spark Shell</td><td>Create Pair RDD</td></tr><tr><td>Create RDD from Collection</td><td>reduceByKey</td></tr></table></td><td></td><td></td><td>CO5</td></tr>
</table>

	Read Text File into RDD (from HDFS)	groupByKey — With Output Viewing			
	Map Transformation	mapValues			
	Filter Transformation	sortByKey			
	Reduce Action	join			
	Collect Action	cogroup			
	Save RDD to HDFS	aggregateByKey			
	Cache/Persist RDD	foldByKey			
		Read from Local File			
7	Visualization using Tableau: Tableau: Tool Overview, Importing Data, Analyzing with Charts, Creating Dashboards, working with maps				CO6

1. HDFS: List of Commands (mkdir, touchz, copy from local/put, copy to local/get, move from local, cp, rmr, du, dus, stat)

Installation of Oracle Virtual Machine and
ClouderaQuickstart

Steps:

Step 1: Download oracle virtual machine and install it as run as administrator

Step 2: Download Cloudera-quickstart-vm-5.4.2-0-Virtualbox and unzip it use 7 zip.

Step 3: Import Cloudera-quickstart-vm-5.4.2-0-Virtualbox.ovf in oracle Virtual Machine.

Step 4: Provide the Memory and processor to machine. It requires minimum 4 gb ram and 2 processor.

Step 5: Enable Network Adapter and attach it to NAT.

Step 6: Start the Machine.

Step 7: Launch Cloudera express. After launch we get command i.e sudo /home/cloudera-manager - force We have to run this command by adding - - express command.

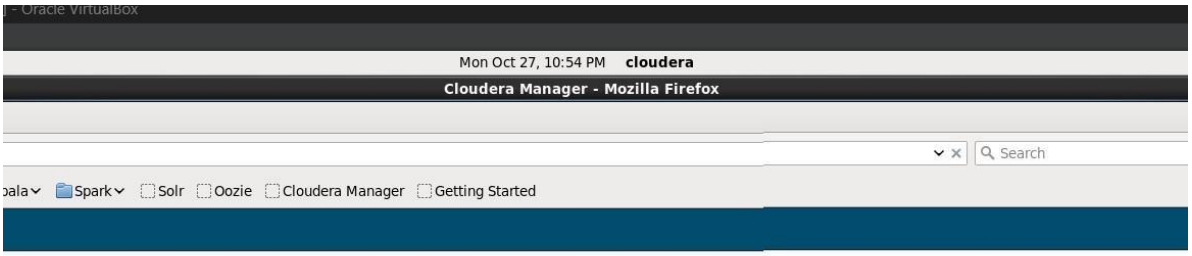
Step 8: we have to login the cloudera manager and after that we have to start the services of Hadoop system.

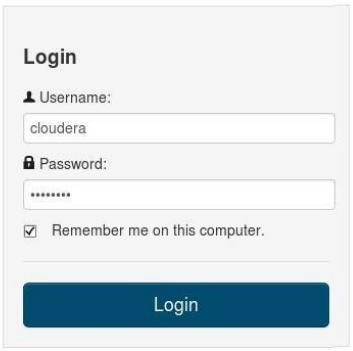
BIGDATA

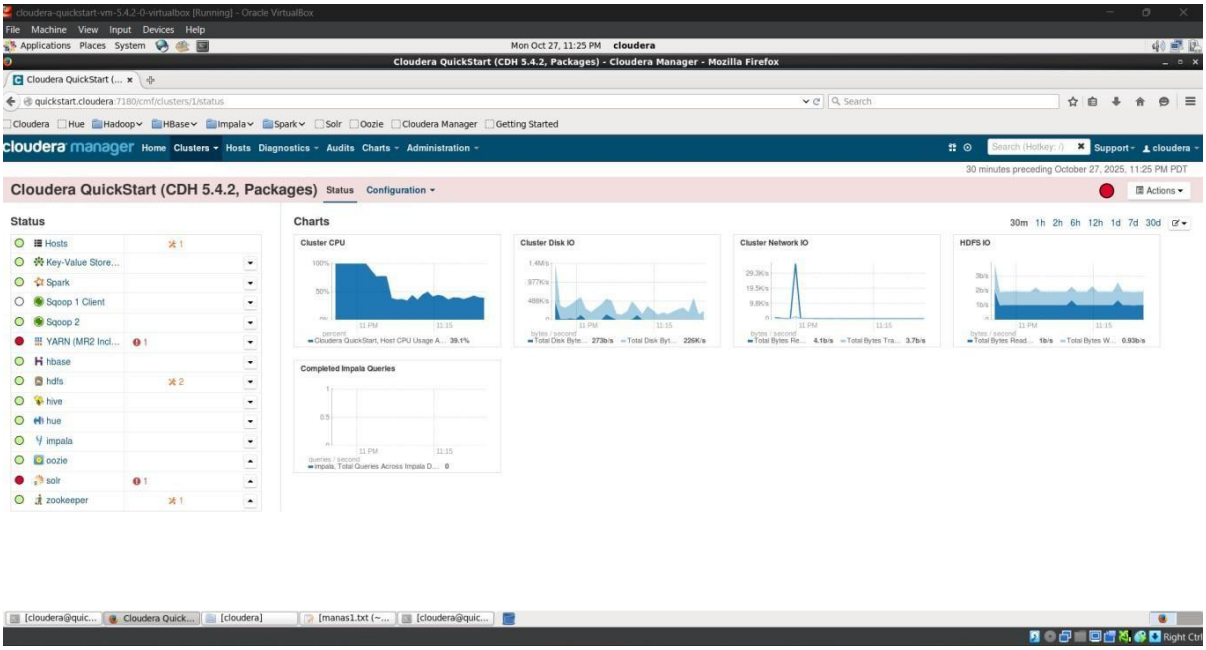

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Mon Oct 27, 10:50 PM cloudera
Computer
cloudera's Home
Eclipse
Launch Cloudera Express
Trash
cloudera
Ask Bigger Questions


cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ sudo /home/cloudera/cloudera-manager --force --express
[QuickStart] Shutting down CDH services via init scripts...
[QuickStart] Disabling CDH services on boot...
[QuickStart] Starting Cloudera Manager daemons...
[QuickStart] Waiting for Cloudera Manager API...
[QuickStart] Configuring deployment...
[QuickStart] Deploying client configuration...
[QuickStart] Starting Cloudera Management Service...
[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 ~]$

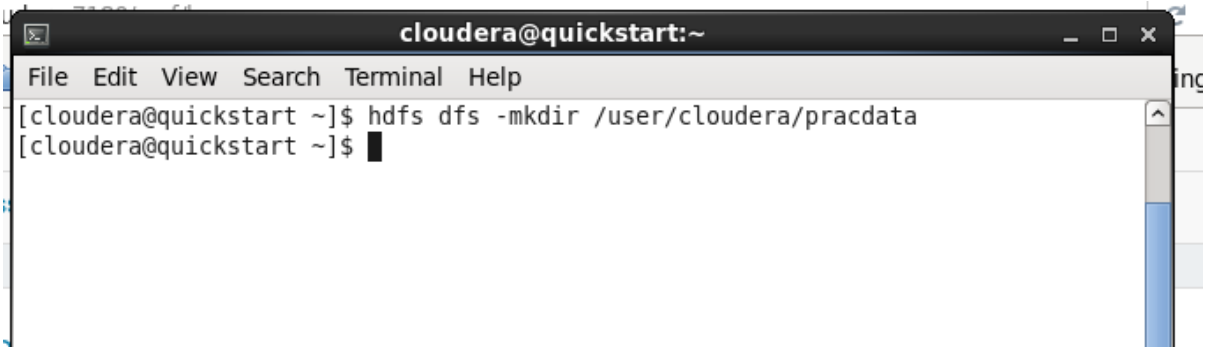
BIGDATA


Mon Oct 27, 10:54 PM cloudera
Cloudera Manager - Mozilla Firefox
Search
Spark
Solr
Oozie
Cloudera Manager
Getting Started

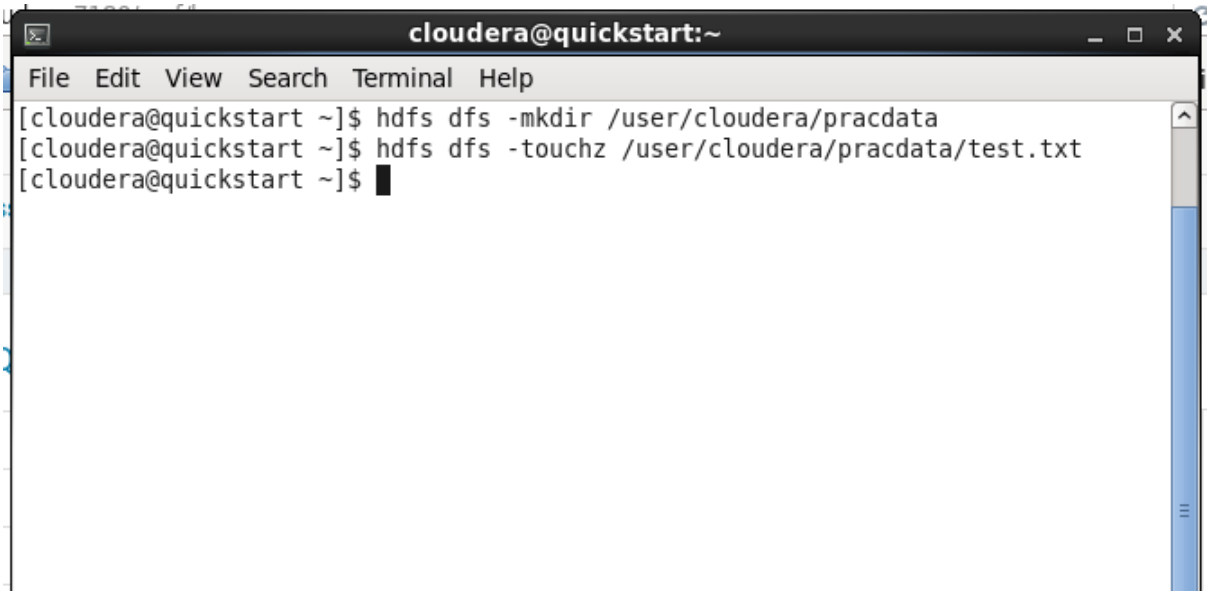

Login
Username:
cloudera
Password:
Remember me on this computer.
Login


Cloudera QuickStart (CDH 5.4.2, Packages) - Cloudera Manager - Mozilla Firefox
Cloudera QuickStart (CDH 5.4.2, Packages)
Status
Charts
Cluster CPU
Cluster Disk IO
Cluster Network IO
HDFS IO
Completed Impala Queries

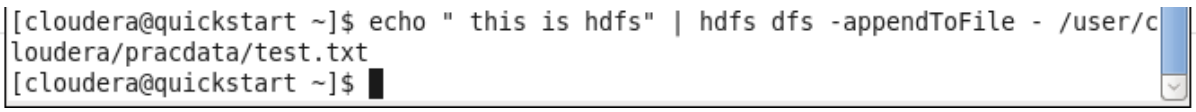
-mkdir
Create directories in HDFS



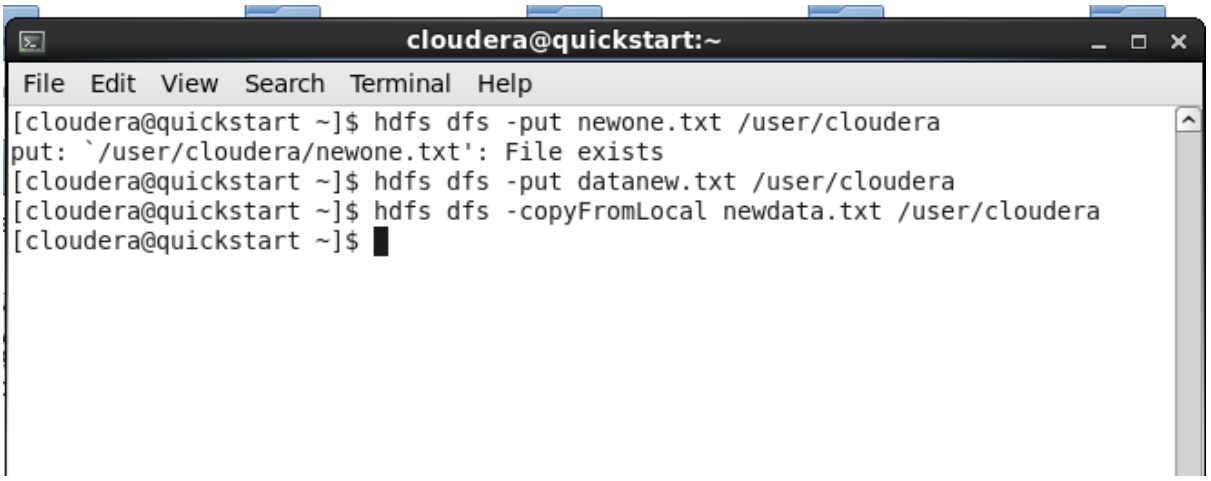
-touchz
Create an empty file in HDFS



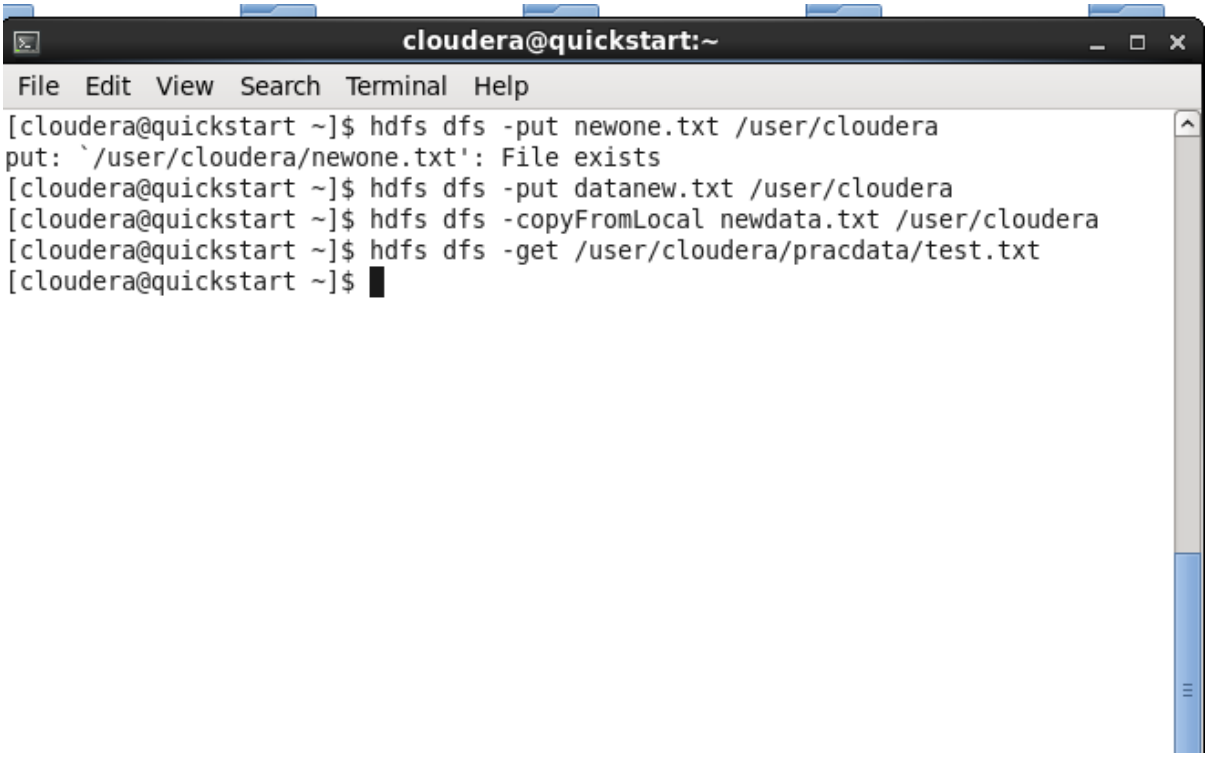
-appendToFile
Add the content to the file



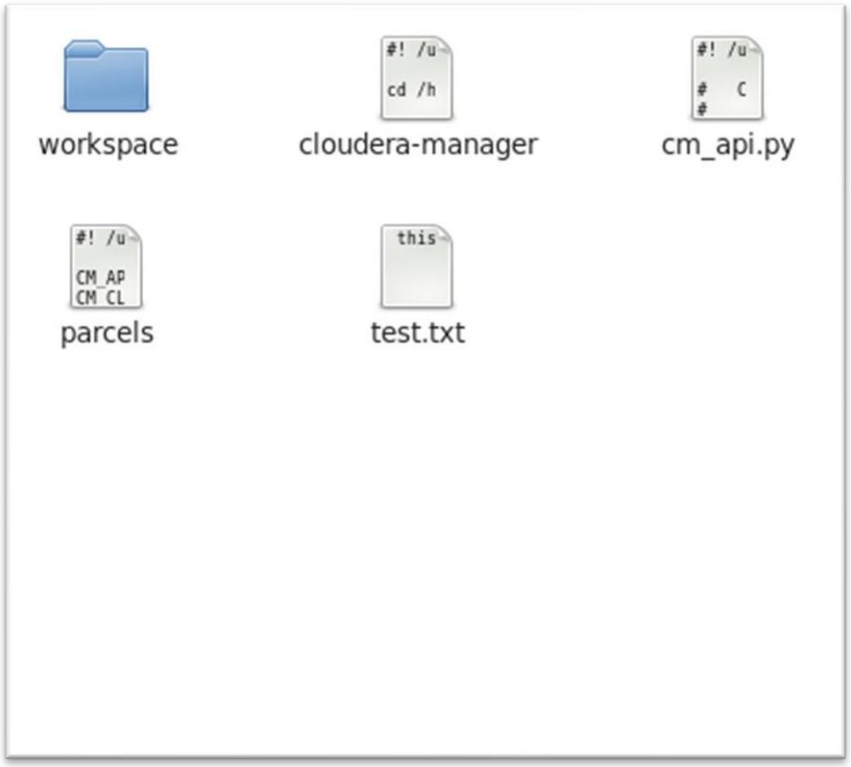
-put or -copyFromLocal
Copy file from local to HDFS



-get or -copyToLocal
Copy file from HDFS to local

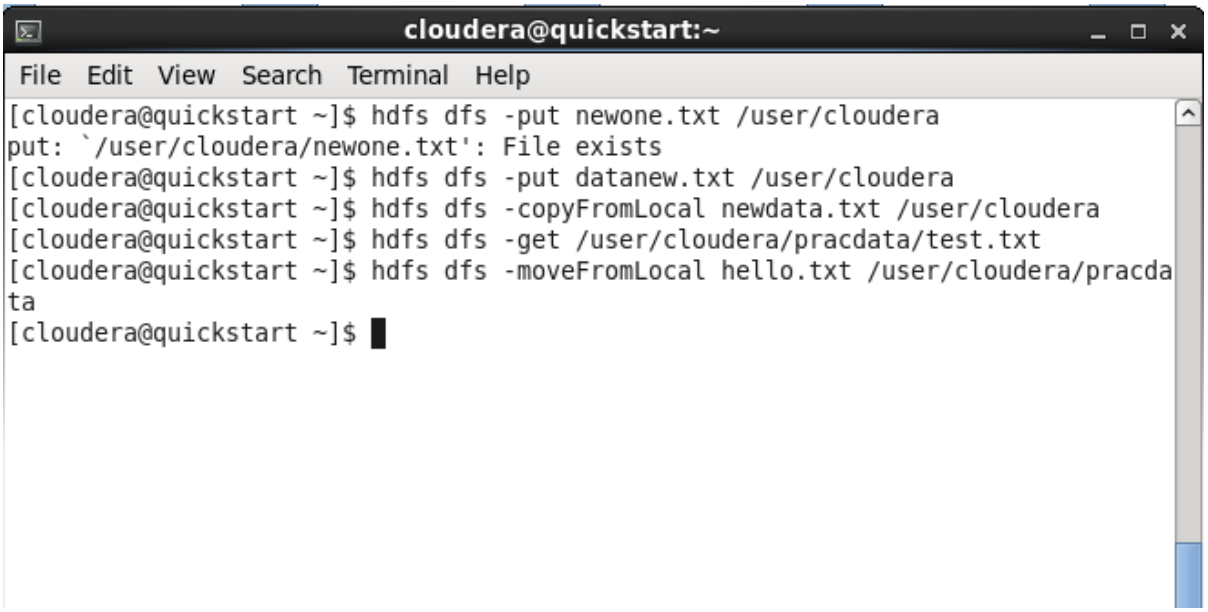


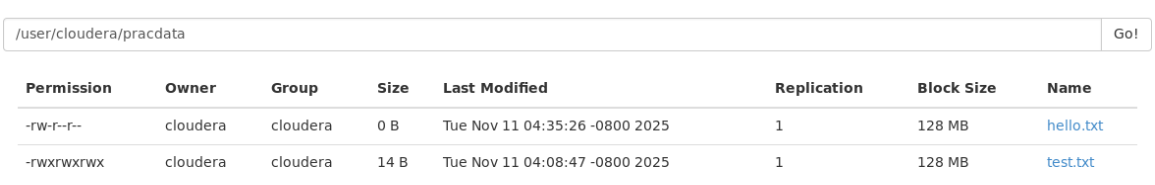
BIGDATA



-moveFromLocal

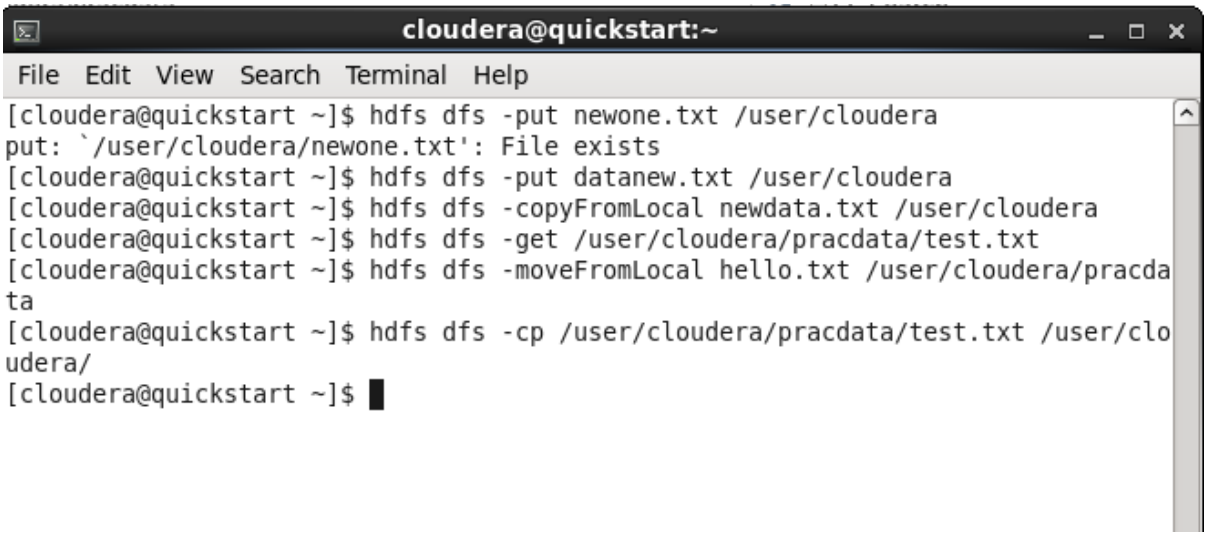
Move (not copy) file from local to HDFS

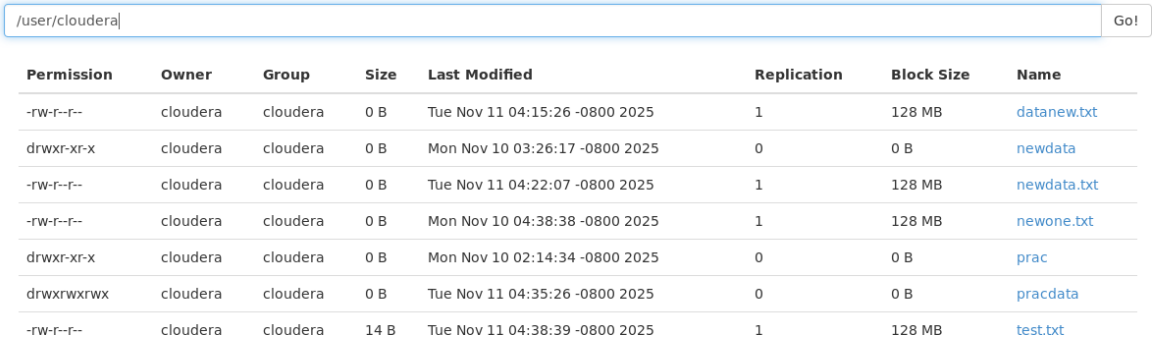




-cp

Copy file within HDFS

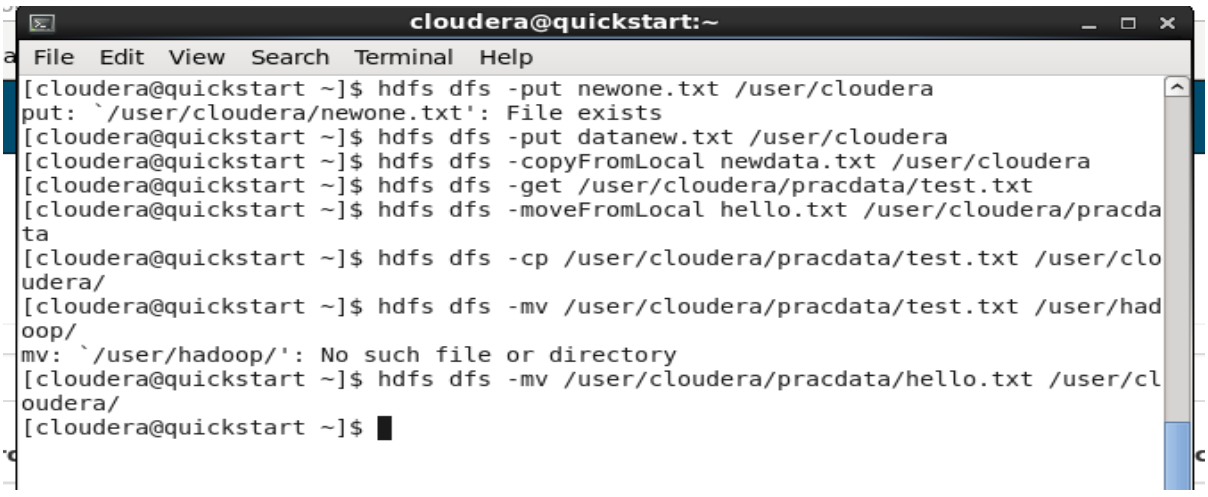




Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	cloudera	cloudera	0 B	Tue Nov 11 04:15:26 -0800 2025	1	128 MB	datanew.txt
drwxr-xr-x	cloudera	cloudera	0 B	Mon Nov 10 03:26:17 -0800 2025	0	0 B	newdata
-rw-r--r--	cloudera	cloudera	0 B	Tue Nov 11 04:22:07 -0800 2025	1	128 MB	newdata.txt
-rw-r--r--	cloudera	cloudera	0 B	Mon Nov 10 04:38:38 -0800 2025	1	128 MB	newone.txt
drwxr-xr-x	cloudera	cloudera	0 B	Mon Nov 10 02:14:34 -0800 2025	0	0 B	prac
drwxrwxrwx	cloudera	cloudera	0 B	Tue Nov 11 04:35:26 -0800 2025	0	0 B	pracdata
-rw-r--r--	cloudera	cloudera	14 B	Tue Nov 11 04:38:39 -0800 2025	1	128 MB	test.txt

-mv

Move file within HDFS



BIGDATA

Browse Directory

/user/cloudera/pracdata Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	cloudera	cloudera	0 B	Tue Nov 11 04:35:26 -0800 2025	1	128 MB	hello.txt
-rwxrwxrwx	cloudera	cloudera	14 B	Tue Nov 11 04:08:47 -0800 2025	1	128 MB	test.txt

Browse Directory

/user/cloudera/pracdata Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rwxrwxrwx	cloudera	cloudera	14 B	Tue Nov 11 04:08:47 -0800 2025	1	128 MB	test.txt

Browse Directory

/user/cloudera/ Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	cloudera	cloudera	0 B	Tue Nov 11 04:15:26 -0800 2025	1	128 MB	datanew.txt
-rw-r--r--	cloudera	cloudera	0 B	Tue Nov 11 04:35:26 -0800 2025	1	128 MB	hello.txt
drwxr-xr-x	cloudera	cloudera	0 B	Mon Nov 10 03:26:17 -0800 2025	0	0 B	newdata

-rm

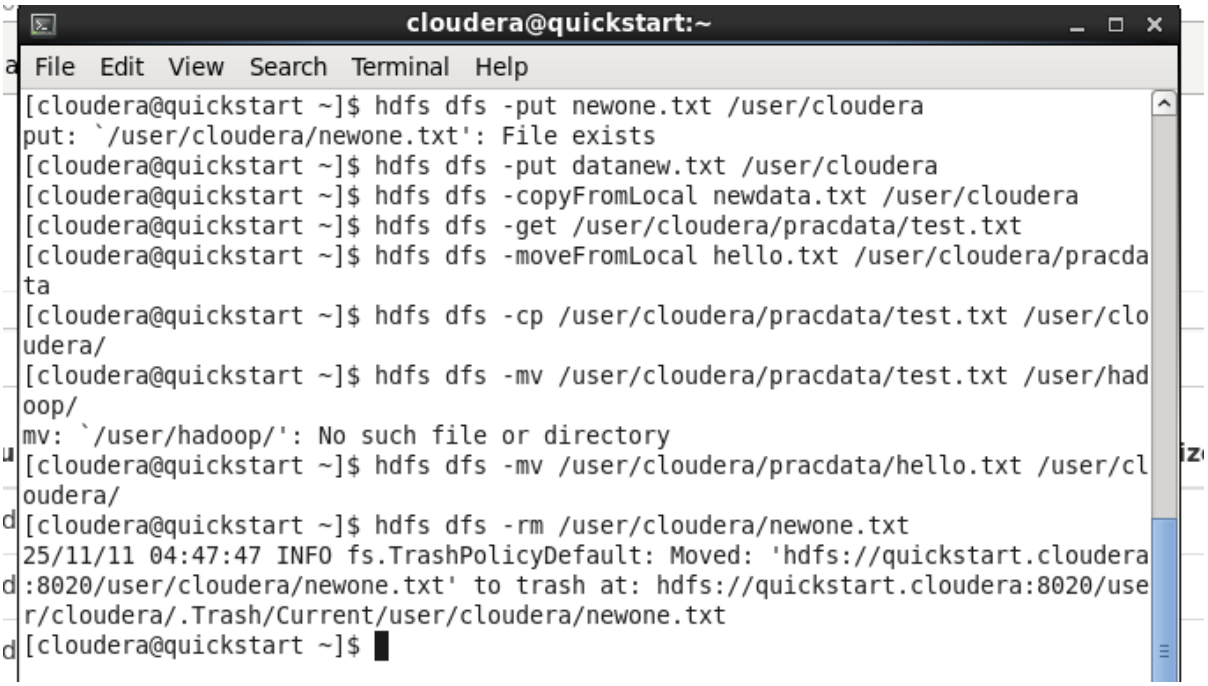
Delete a file

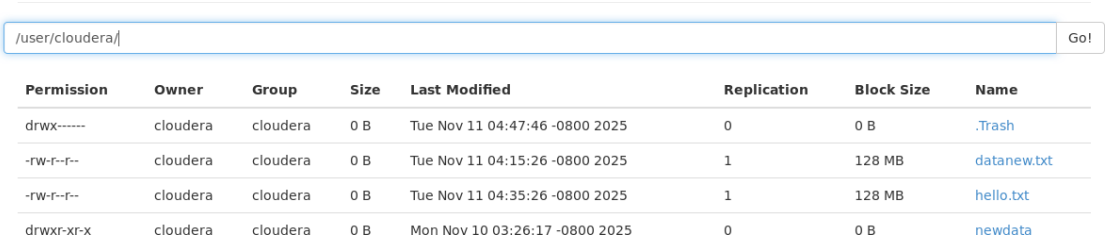
Browse Directory

/user/cloudera/ Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	cloudera	cloudera	0 B	Tue Nov 11 04:15:26 -0800 2025	1	128 MB	datanew.txt
-rw-r--r--	cloudera	cloudera	0 B	Tue Nov 11 04:35:26 -0800 2025	1	128 MB	hello.txt
drwxr-xr-x	cloudera	cloudera	0 B	Mon Nov 10 03:26:17 -0800 2025	0	0 B	newdata
-rw-r--r--	cloudera	cloudera	0 B	Tue Nov 11 04:22:07 -0800 2025	1	128 MB	newdata.txt
-rw-r--r--	cloudera	cloudera	0 B	Mon Nov 10 04:38:38 -0800 2025	1	128 MB	newone.txt

BIGDATA

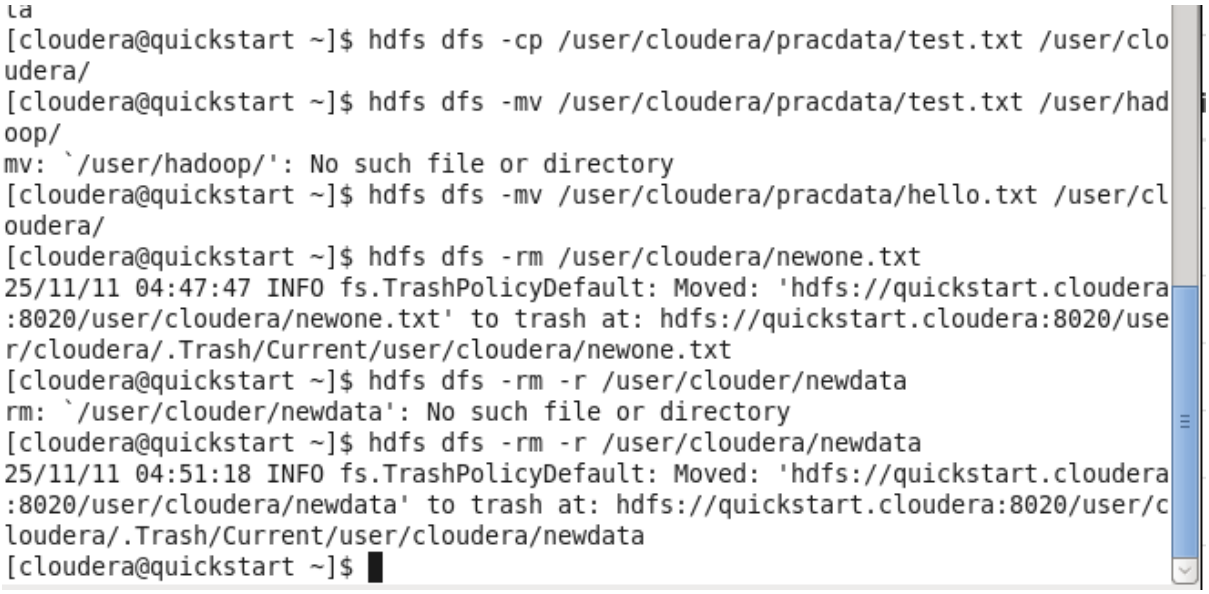




Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwx------	cloudera	cloudera	0 B	Tue Nov 11 04:47:46 -0800 2025	0	0 B	.Trash
-rw-r--r--	cloudera	cloudera	0 B	Tue Nov 11 04:15:26 -0800 2025	1	128 MB	datanew.txt
-rw-r--r--	cloudera	cloudera	0 B	Tue Nov 11 04:35:26 -0800 2025	1	128 MB	hello.txt
drwxr-xr-x	cloudera	cloudera	0 B	Mon Nov 10 03:26:17 -0800 2025	0	0 B	newdata

-rm -r or -rm –R

Delete a directory recursively



BIGDATA

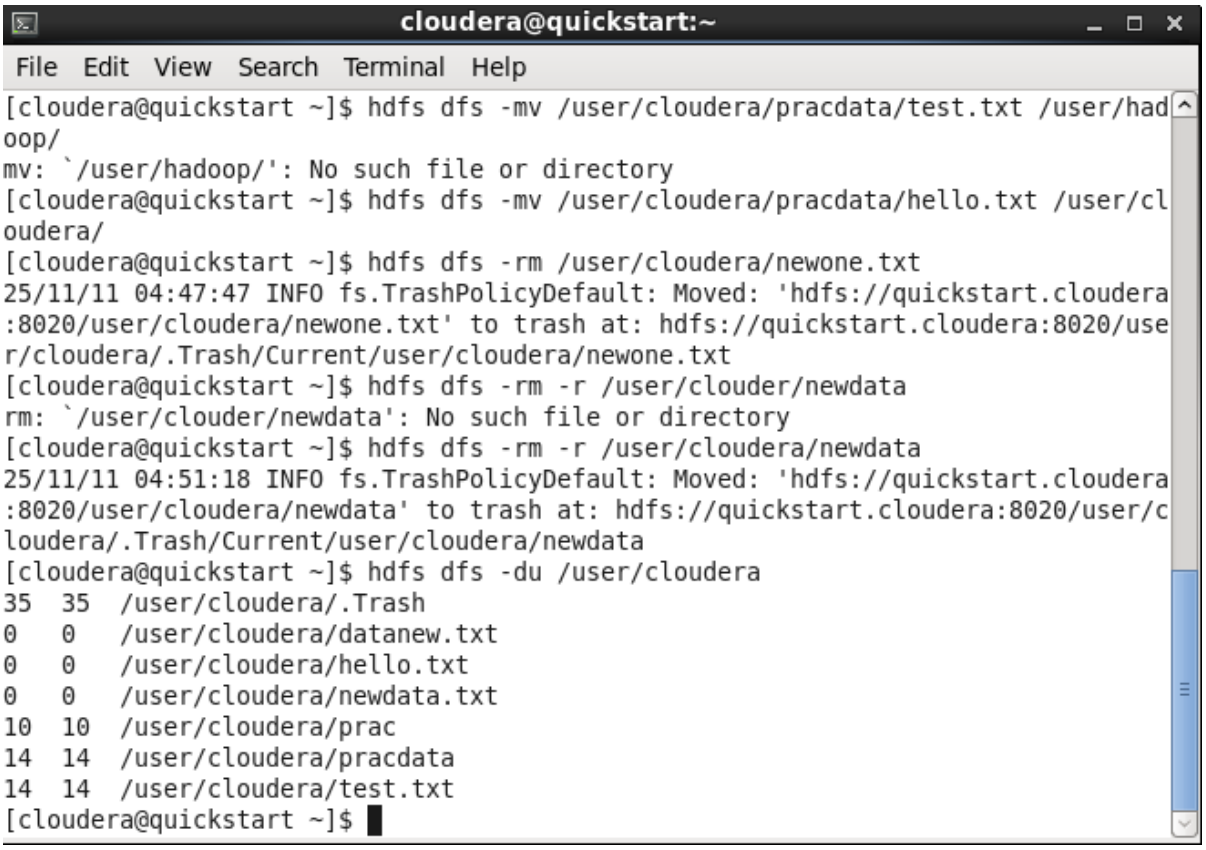
Browse Directory

/user/cloudera/ Go!

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
drwx------	cloudera	cloudera	0 B	Tue Nov 11 04:47:46 -0800 2025	0	0 B	.Trash
-rw-r--r--	cloudera	cloudera	0 B	Tue Nov 11 04:15:26 -0800 2025	1	128 MB	datanew.txt
-rw-r--r--	cloudera	cloudera	0 B	Tue Nov 11 04:35:26 -0800 2025	1	128 MB	hello.txt
drwxr-xr-x	cloudera	cloudera	0 B	Mon Nov 10 03:26:17 -0800 2025	0	0 B	newdata
-rw-r--r--	cloudera	cloudera	0 B	Tue Nov 11 04:22:07 -0800 2025	1	128 MB	newdata.txt

-du

Show file/directory size



-dus

Show total size summary

```
35  35  /user/cloudera/.Trash
0   0   /user/cloudera/datanew.txt
0   0   /user/cloudera/hello.txt
0   0   /user/cloudera/newdata.txt
10  10  /user/cloudera/prac
14  14  /user/cloudera/pracdata
14  14  /user/cloudera/test.txt
[cloudera@quickstart ~]$ hdfs dfs -dus /user/cloudera
dus: DEPRECATED: Please use 'du -s' instead.
73  73  /user/cloudera
[cloudera@quickstart ~]$ hdfs dfs -du -s /user/cloudera
73  73  /user/cloudera
[cloudera@quickstart ~]$
```

-stat

Display file metadata

```
[cloudera@quickstart ~]$ hdfs dfs -du -s /user/cloudera
73  73  /user/cloudera
[cloudera@quickstart ~]$ hdfs dfs -stat /user/cloudera/pracdata/test.txt
2025-11-11 12:08:47
[cloudera@quickstart ~]$
```

2. Map Reduce

A. Write a program in Map Reduce for Word Count operation.

Step 1: Open virtual box and then start cloudera quickstart

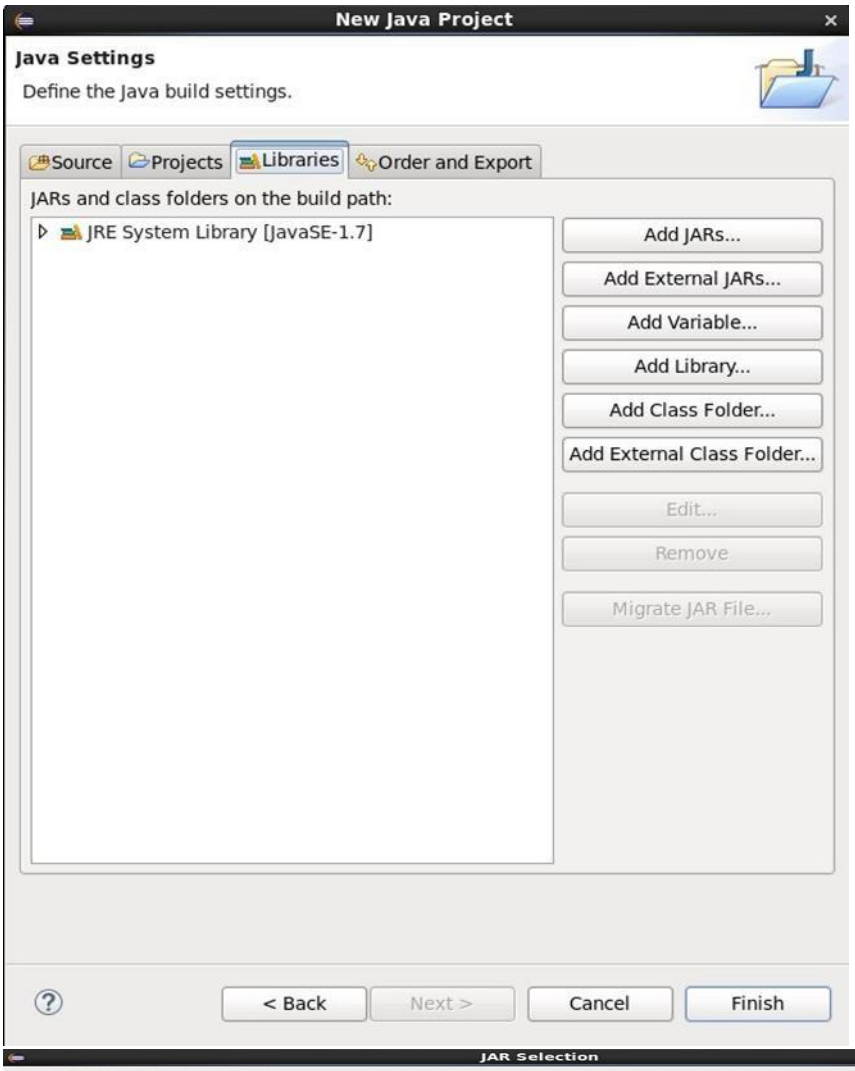
Step 2: Open eclipse present on the cloudera desktop

Step 3: Create java project

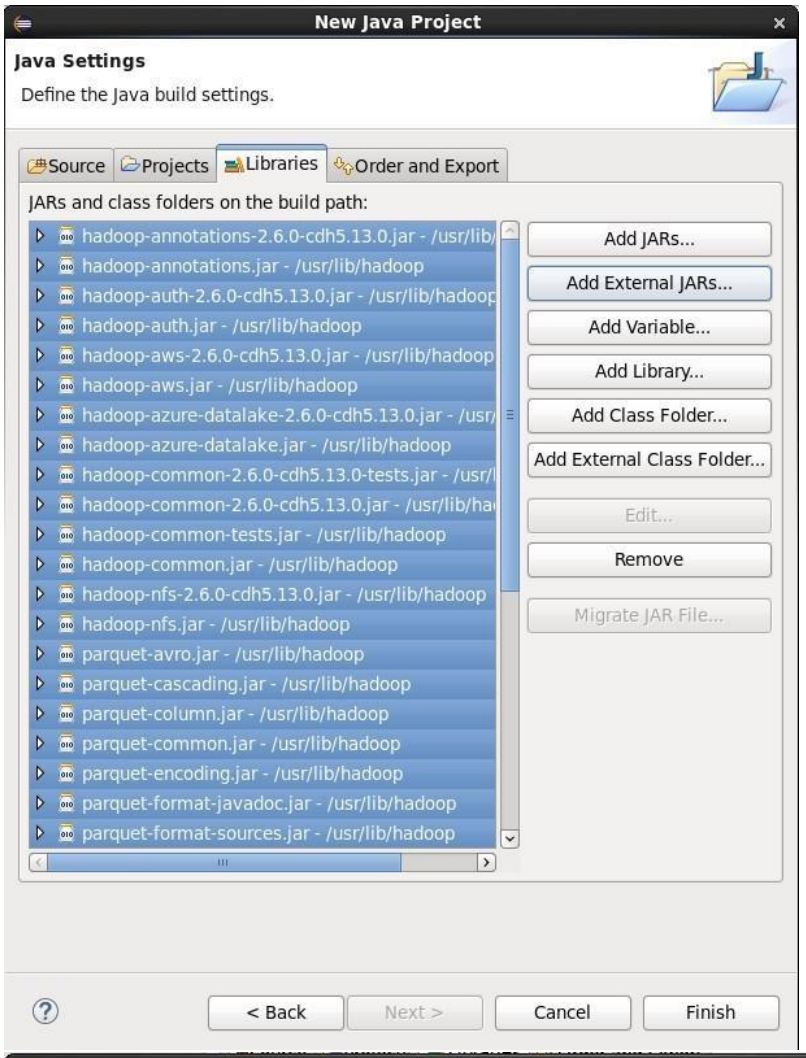
- File->New-> Java Project
- Give project name: Practical_4
- Click Next

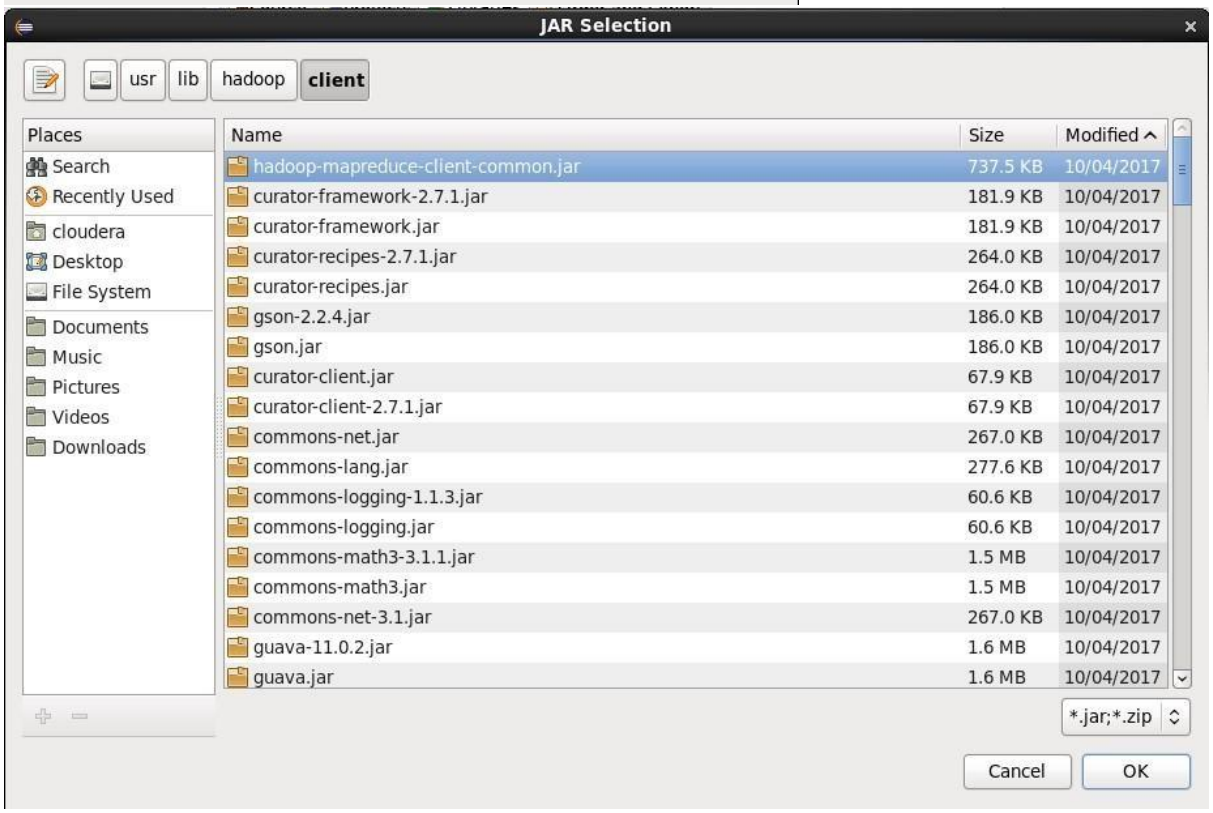
Step 4: Add Hadoop libraries to project

- Select Libraries tab->click on Add External Jars
- File System->user->lib->Hadoop
- Select all library(jar) files->ok
- Again click on Add External Jars
- File System->user->lib->Hadoop->client
- Select all library(jar) files from client->ok->finish



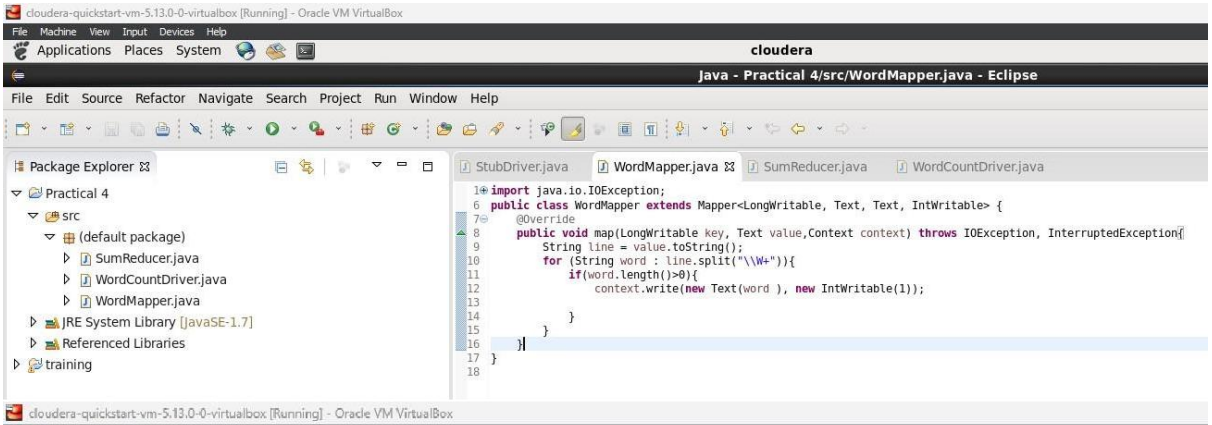
BIGDATA

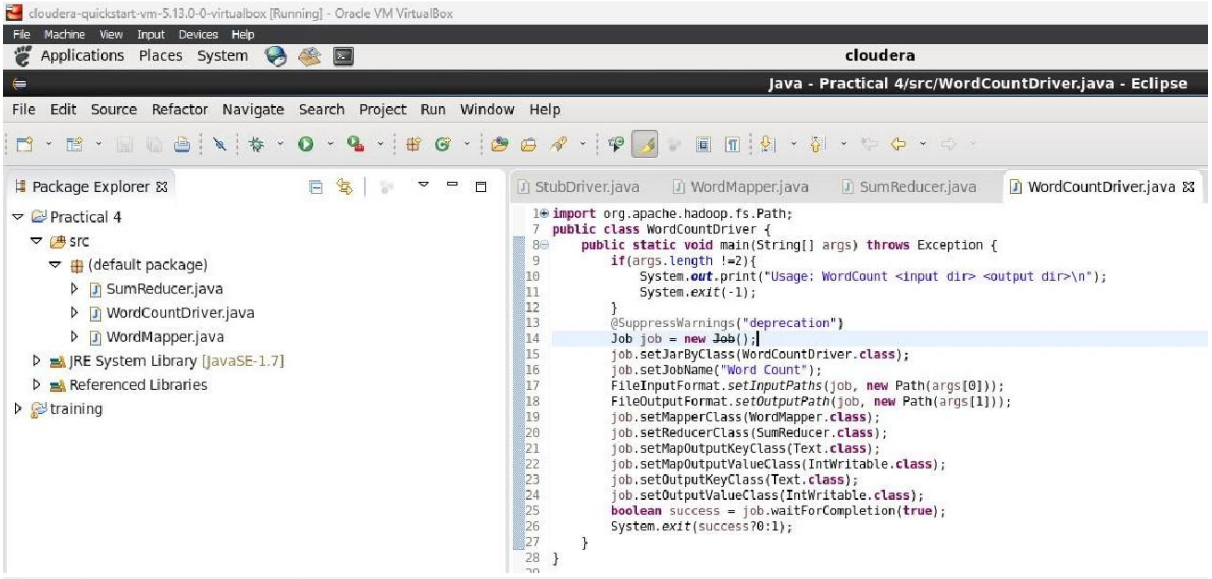

New Java Project
Java Settings
Define the Java build settings.
Source
Projects
Libraries
Order and Export
JARs and class folders on the build path:
hadoop-annotations-2.6.0-cdh5.13.0.jar - /usr/lib/
hadoop-annotations.jar - /usr/lib/hadoop
hadoop-auth-2.6.0-cdh5.13.0.jar - /usr/lib/hadoop
hadoop-auth.jar - /usr/lib/hadoop
hadoop-aws-2.6.0-cdh5.13.0.jar - /usr/lib/hadoop
hadoop-aws.jar - /usr/lib/hadoop
hadoop-azure-datalake-2.6.0-cdh5.13.0.jar - /usr/
hadoop-azure-datalake.jar - /usr/lib/hadoop
hadoop-common-2.6.0-cdh5.13.0-tests.jar - /usr/
hadoop-common-2.6.0-cdh5.13.0.jar - /usr/lib/ha
hadoop-common-tests.jar - /usr/lib/hadoop
hadoop-common.jar - /usr/lib/hadoop
hadoop-nfs-2.6.0-cdh5.13.0.jar - /usr/lib/hadoop
hadoop-nfs.jar - /usr/lib/hadoop
parquet-avro.jar - /usr/lib/hadoop
parquet-cascading.jar - /usr/lib/hadoop
parquet-column.jar - /usr/lib/hadoop
parquet-common.jar - /usr/lib/hadoop
parquet-encoding.jar - /usr/lib/hadoop
parquet-format-javadoc.jar - /usr/lib/hadoop
parquet-format-sources.jar - /usr/lib/hadoop
Add JARs...
Add External JARs...
Add Variable...
Add Library...
Add Class Folder...
Add External Class Folder...
Edit...
Remove
Migrate JAR File...
< Back
Next >
Cancel
Finish

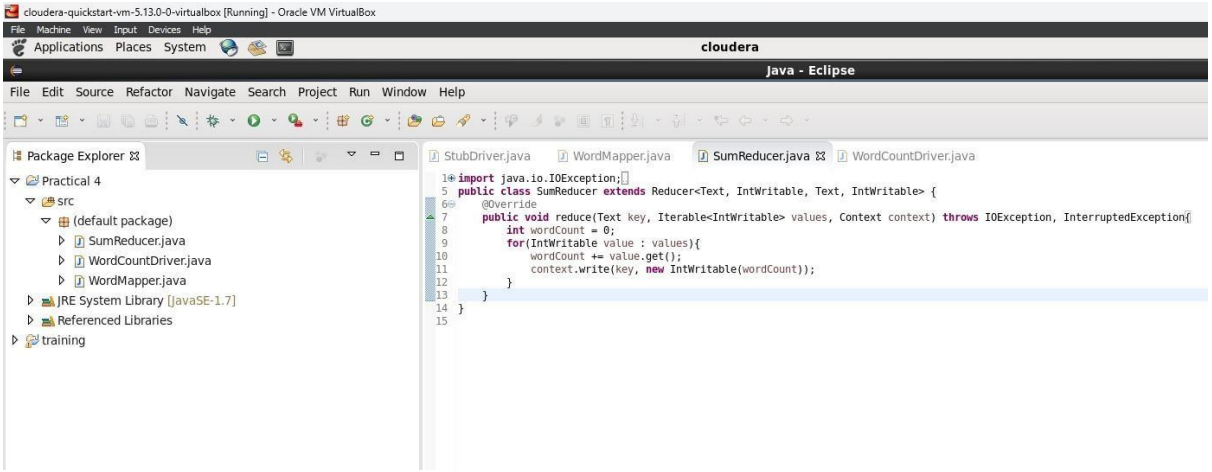

JAR Selection
usr
lib
hadoop
client
Places
Search
Recently Used
cloudera
Desktop
File System
Documents
Music
Pictures
Videos
Downloads
Name
Size
Modified
hadoop-mapreduce-client-common.jar 737.5 KB 10/04/2017
curator-framework-2.7.1.jar 181.9 KB 10/04/2017
curator-framework.jar 181.9 KB 10/04/2017
curator-recipes-2.7.1.jar 264.0 KB 10/04/2017
curator-recipes.jar 264.0 KB 10/04/2017
gson-2.2.4.jar 186.0 KB 10/04/2017
gson.jar 186.0 KB 10/04/2017
curator-client.jar 67.9 KB 10/04/2017
curator-client-2.7.1.jar 67.9 KB 10/04/2017
commons-net.jar 267.0 KB 10/04/2017
commons-lang.jar 277.6 KB 10/04/2017
commons-logging-1.1.3.jar 60.6 KB 10/04/2017
commons-logging.jar 60.6 KB 10/04/2017
commons-math3-3.1.1.jar 1.5 MB 10/04/2017
commons-math3.jar 1.5 MB 10/04/2017
commons-net-3.1.jar 267.0 KB 10/04/2017
guava-11.0.2.jar 1.6 MB 10/04/2017
guava.jar 1.6 MB 10/04/2017
.jar;.zip
Cancel
OK

Step 5: Write java code for word count

- Right click on src folder of project WordCount
- New->class
- Write class name WordMapper
- Click Finish
- Write the code for WordMapper
- Same way create classes SumReducer and WordCountDriver

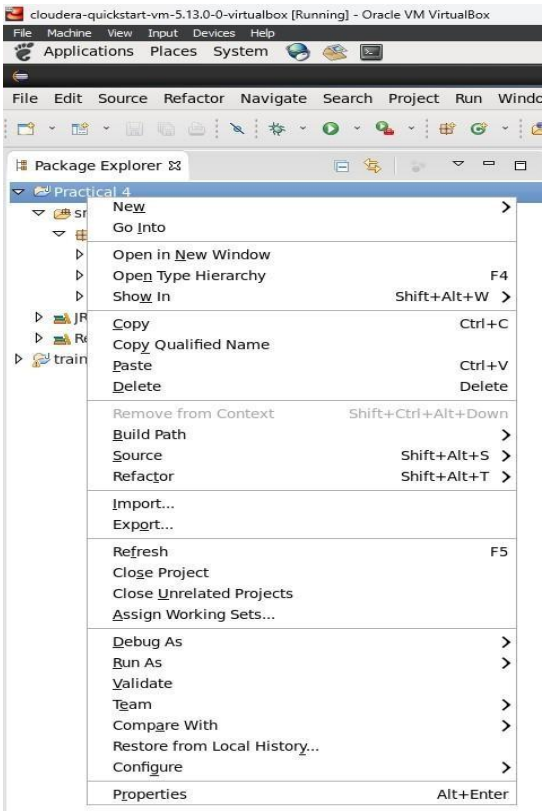


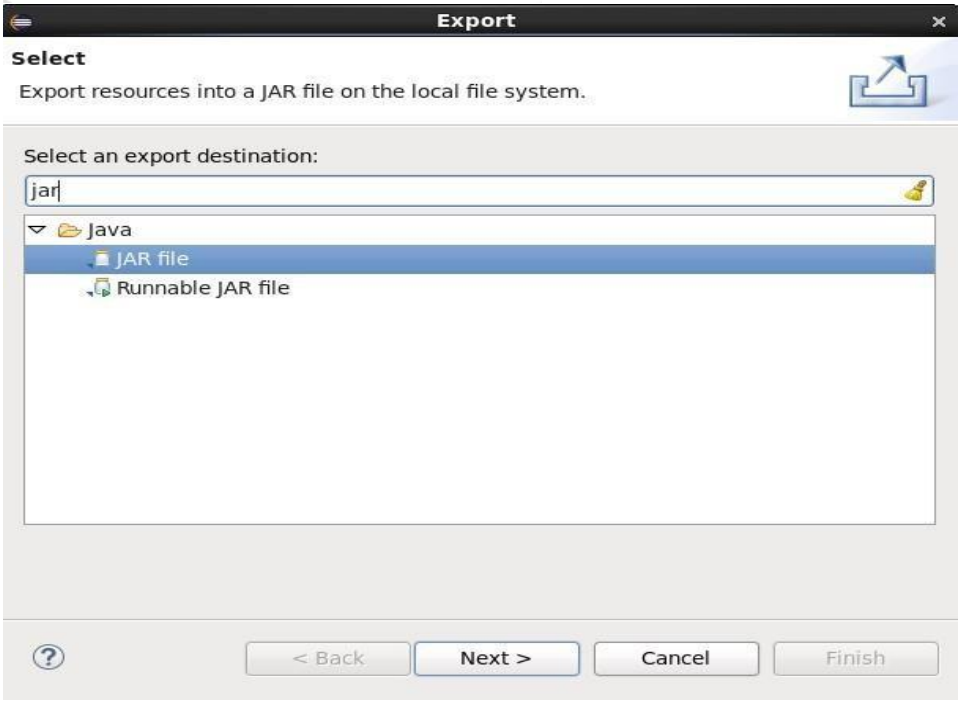


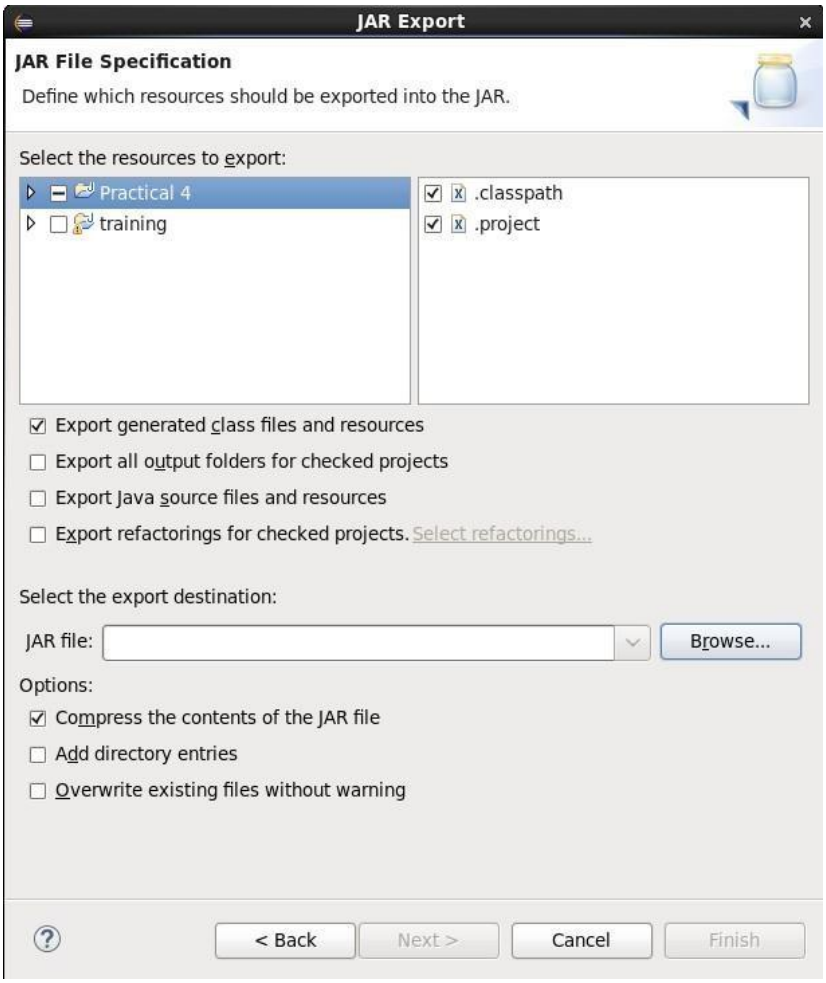


Step 6: Export the project as jar

- Right click on project Practical 4 and select Export>> Java>>JAR file>>Next
- Select the export destination-Click browse-give file name wordcount.jar
- Click ok>>Finish>>ok







Verify the jar file

- Verify the jar file through command line open terminal give command
- ls

Step 7: Move the jar file to the Hadoop file system

- hdfs dfs -put wordcount.jar /user/cloudera
- hdfs dfs –ls

Step 8: Create the input file for the MapReduce program

- Command: cat > myInputFile.txt
- Welcome to the NMITD MCA
- I am persuing MCA in NMITD
- Enter data in input file and press enter and ctrl z
- Command: cat myInputFile.txt

```
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 9 items
drwx------   - cloudera cloudera          0 2025-08-06 22:00 .Trash
drwx------   - cloudera cloudera          0 2025-08-06 23:16 .staging
-rw-r--r--   1 cloudera cloudera       6555 2025-08-06 23:08 MatrixMultiplication.jar
-rw-r--r--   1 cloudera cloudera       4887 2025-08-05 23:20 WordCount.jar
drwxr-xr-x   - cloudera cloudera          0 2025-07-30 00:31 manas
drwxr-xr-x   - cloudera cloudera          0 2025-08-06 23:14 matrixInput
drwxr-xr-x   - cloudera cloudera          0 2025-08-06 23:16 matrixOutput
-rw-r--r--   1 cloudera cloudera         52 2025-08-05 23:25 myInputFile.txt
drwxr-xr-x   - cloudera cloudera          0 2025-08-06 21:43 myOutput
[cloudera@quickstart ~]$ cat > myInputFile.txt
Welcome to the NMITD MCA
I am pursuing MCA in NMITD
^Z
[1]+  Stopped                 cat > myInputFile.txt
[cloudera@quickstart ~]$ cat myInputFile.txt
Welcome to the NMITD MCA
I am pursuing MCA in NMITD
[cloudera@quickstart ~]$ hdfs dfs -put myInputFile.txt /user/cloudera
```

Step 9: Move the input file to the Hadoop file system

- hdfs dfs -put myInputFile.txt /user/cloudera
- hdfs dfs -ls
- hdfs dfs –ls / (here / indicates root directory of Hadoop file system(hdfs))

Step 10: Run mapreduce program on Hadoop

- syntax: hadoop jar jarfilename.jar classname inputfilename.txt outputfoldername
- command: hadoop jar wordcount.jar WordCountDriver myInputFile.txt myOutput

```
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 9 items
drwx------   - cloudera cloudera          0 2025-08-06 22:00 .Trash
drwx------   - cloudera cloudera          0 2025-08-06 23:16 .staging
-rw-r--r--   1 cloudera cloudera       6555 2025-08-06 23:08 MatrixMultiplication.jar
-rw-r--r--   1 cloudera cloudera       4887 2025-08-05 23:20 WordCount.jar
drwxr-xr-x   - cloudera cloudera          0 2025-07-30 00:31 manas
drwxr-xr-x   - cloudera cloudera          0 2025-08-06 23:14 matrixInput
drwxr-xr-x   - cloudera cloudera          0 2025-08-06 23:16 matrixOutput
-rw-r--r--   1 cloudera cloudera         52 2025-08-05 23:25 myInputFile.txt
drwxr-xr-x   - cloudera cloudera          0 2025-08-06 21:43 myOutput
h[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 6 items
drwxrwxrwx   - hdfs  supergroup          0 2017-10-23 09:15 /benchmarks
drwxr-xr-x   - hbase supergroup          0 2025-08-06 21:45 /hbase
drwxr-xr-x   - solr  solr                0 2017-10-23 09:18 /solr
drwxrwxrwt   - hdfs  supergroup          0 2025-08-06 21:38 /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 ~]$ hadoop jar WordCount.jar WordCountDriver myInputFile.txt myOutput
```

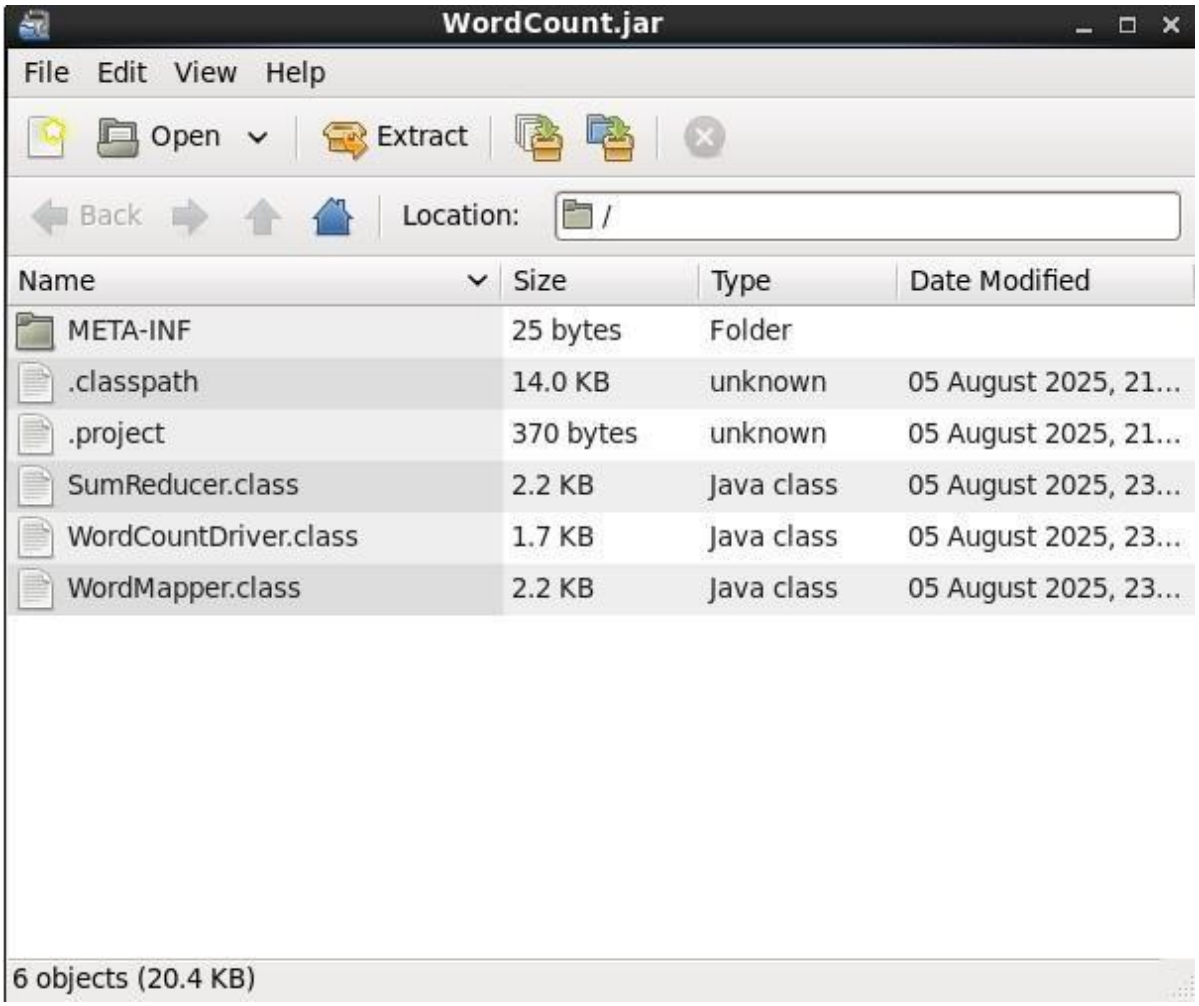
Step 11: view output directory

- hdfs dfs –ls
- hdfs dfs -ls /user/cloudera/myOutput

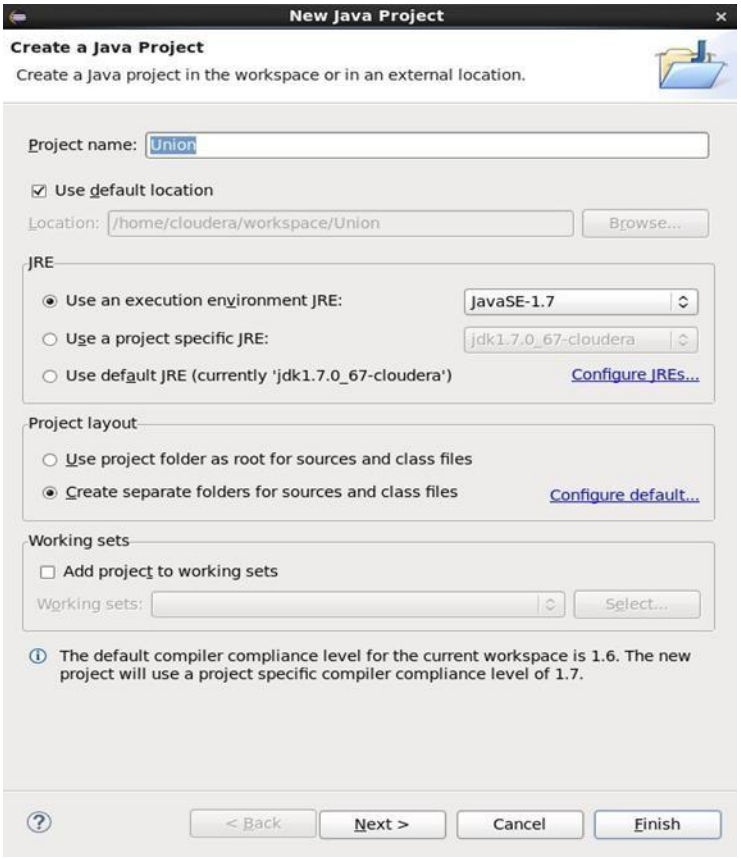
Step 12: view the output file

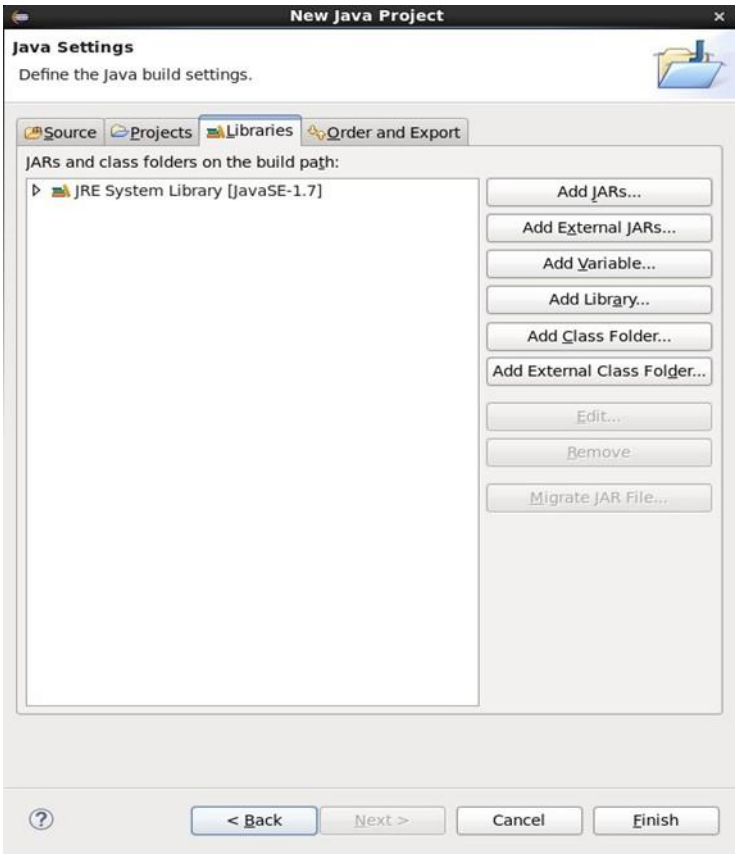
hdfs dfs -cat /user/cloudera/myOutput/part-r-00000

```
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/myOutput
Found 2 items
-rw-r--r--   1 cloudera cloudera          0 2025-08-06 21:43 /user/cloudera/myOutput/_SUCCESS
-rw-r--r--   1 cloudera cloudera         74 2025-08-06 21:43 /user/cloudera/myOutput/part-r-00000
[cloudera@quickstart ~]$ hdfs dfs -cat /user/cloudera/myOutput/part-r-00000
I       1
MCA     1
MCA     2
NMITD   1
NMITD   2
Welcome 1
am      1
in      1
pursuing        1
the     1
to      1
```

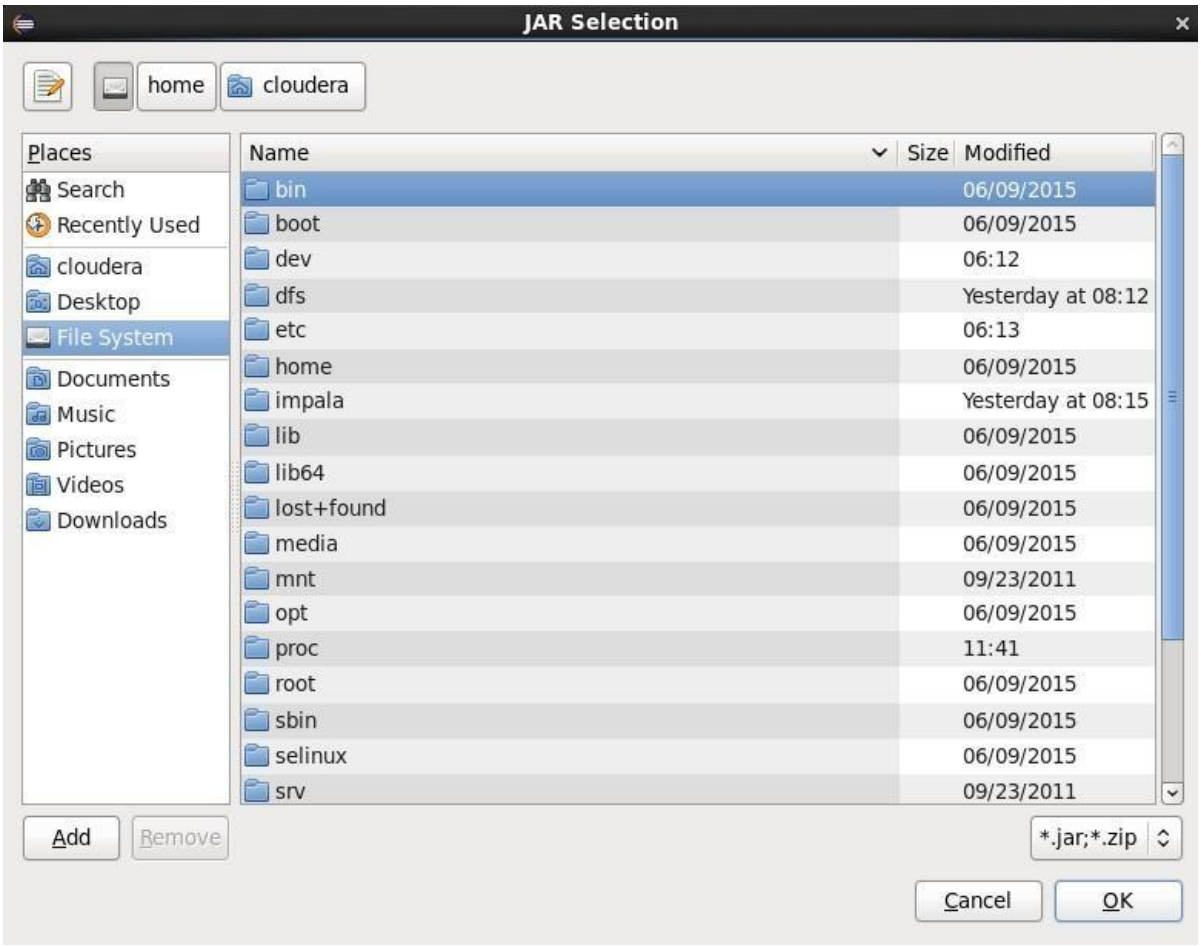


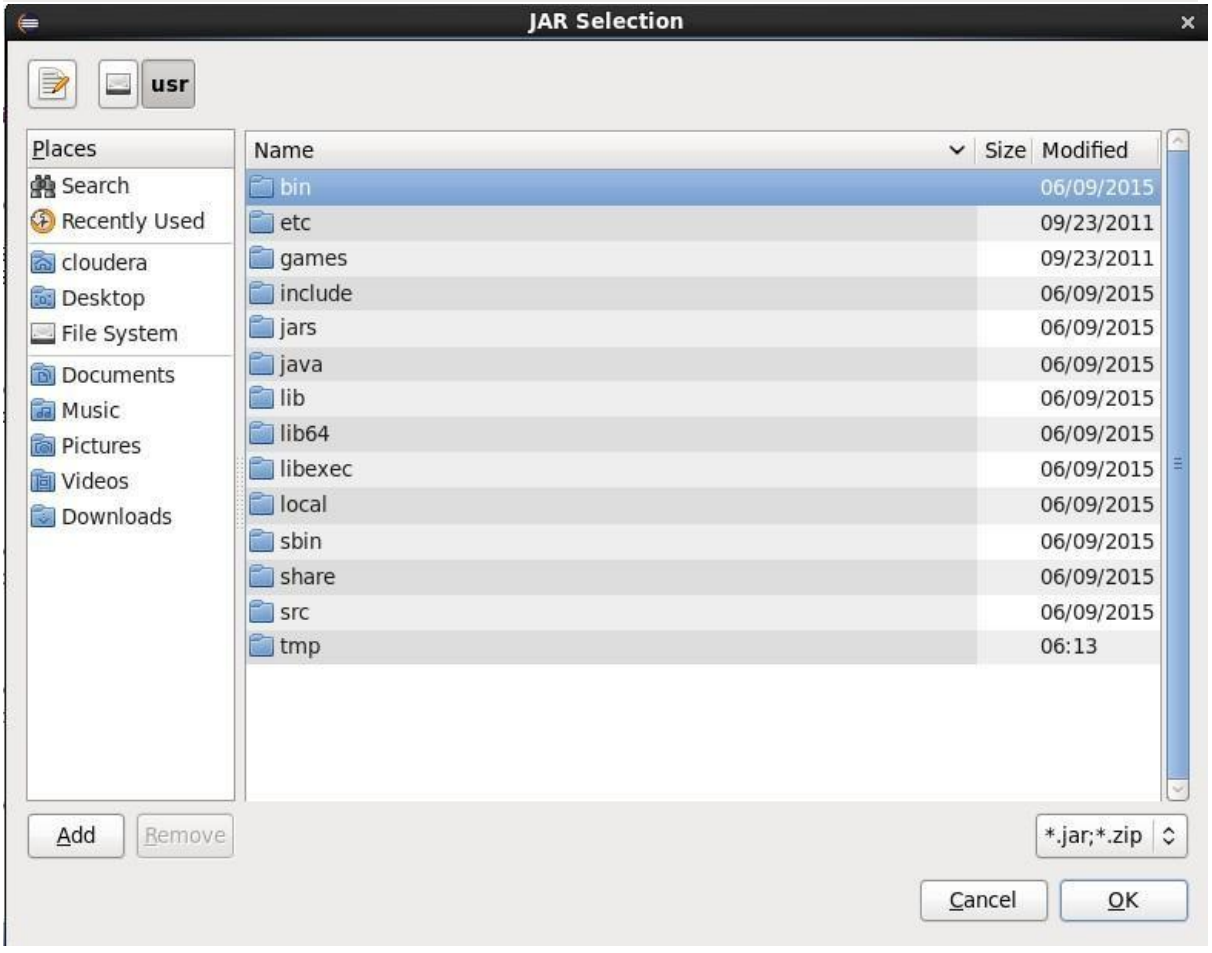
Write a program in Map Reduce for Union operation.



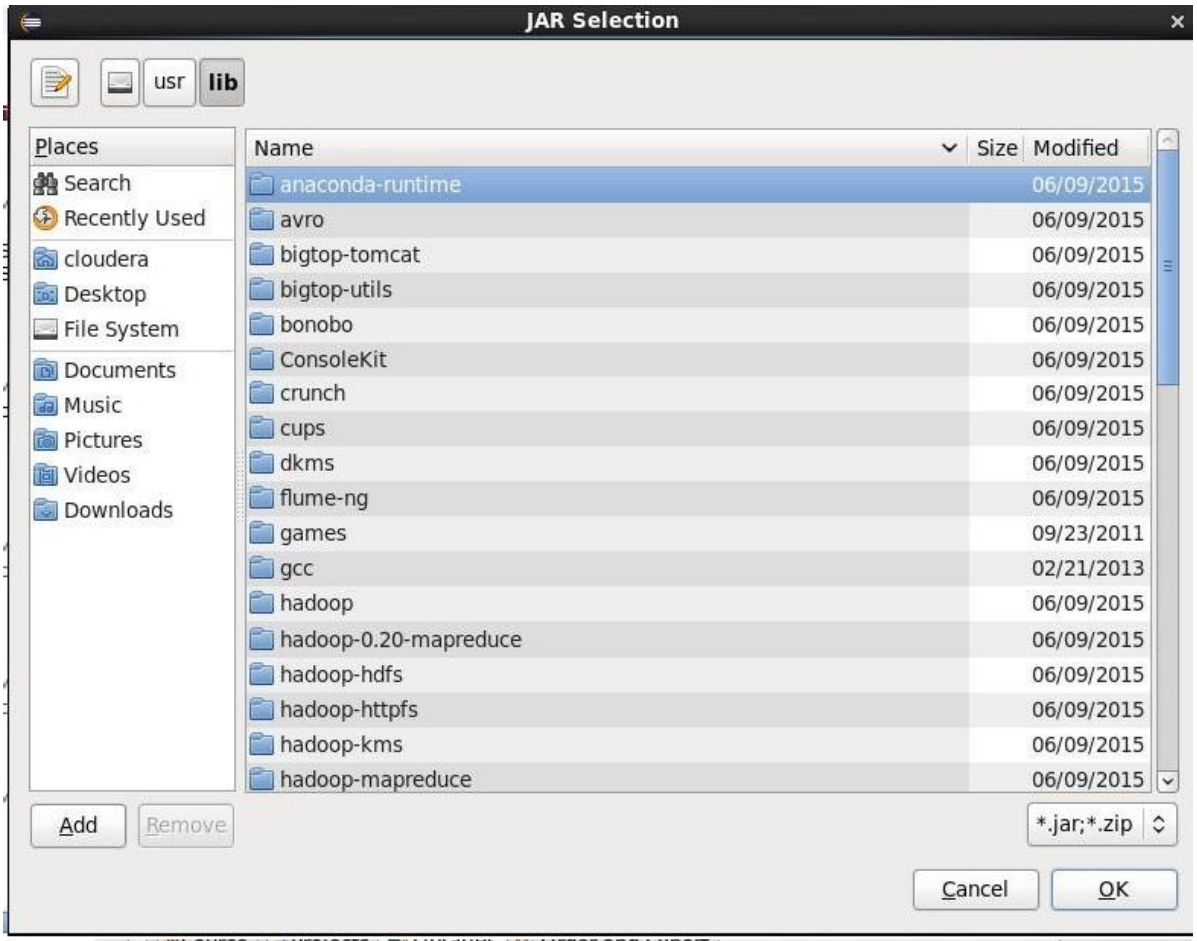


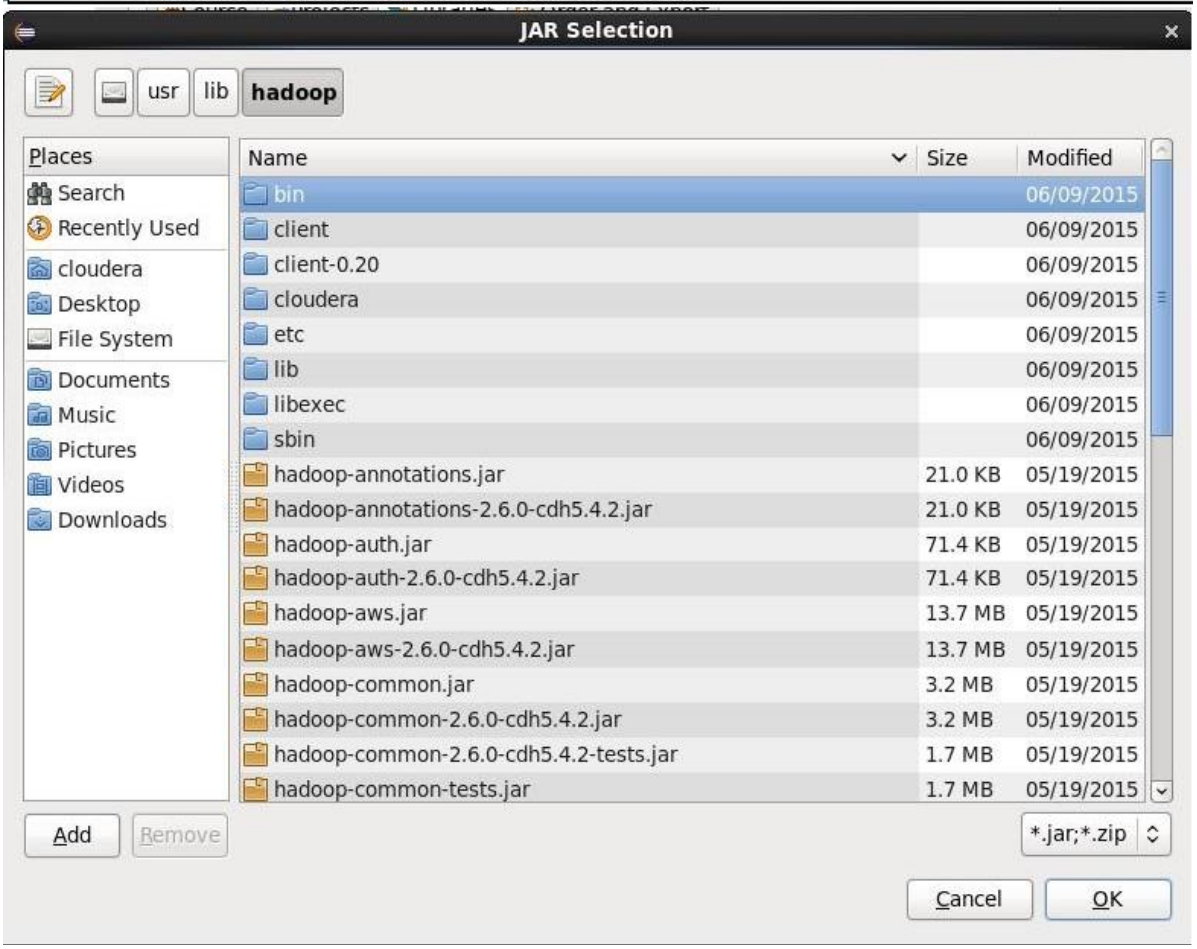
BIGDATA


JAR Selection
home
cloudera
Places
Search
Recently Used
cloudera
Desktop
File System
Documents
Music
Pictures
Videos
Downloads
Name
Size
Modified
bin 06/09/2015
boot 06/09/2015
dev 06:12
dfs Yesterday at 08:12
etc 06:13
home 06/09/2015
impala Yesterday at 08:15
lib 06/09/2015
lib64 06/09/2015
lost+found 06/09/2015
media 06/09/2015
mnt 09/23/2011
opt 06/09/2015
proc 11:41
root 06/09/2015
sbin 06/09/2015
selinux 06/09/2015
srv 09/23/2011
Add
Remove
.jar;.zip
Cancel
OK

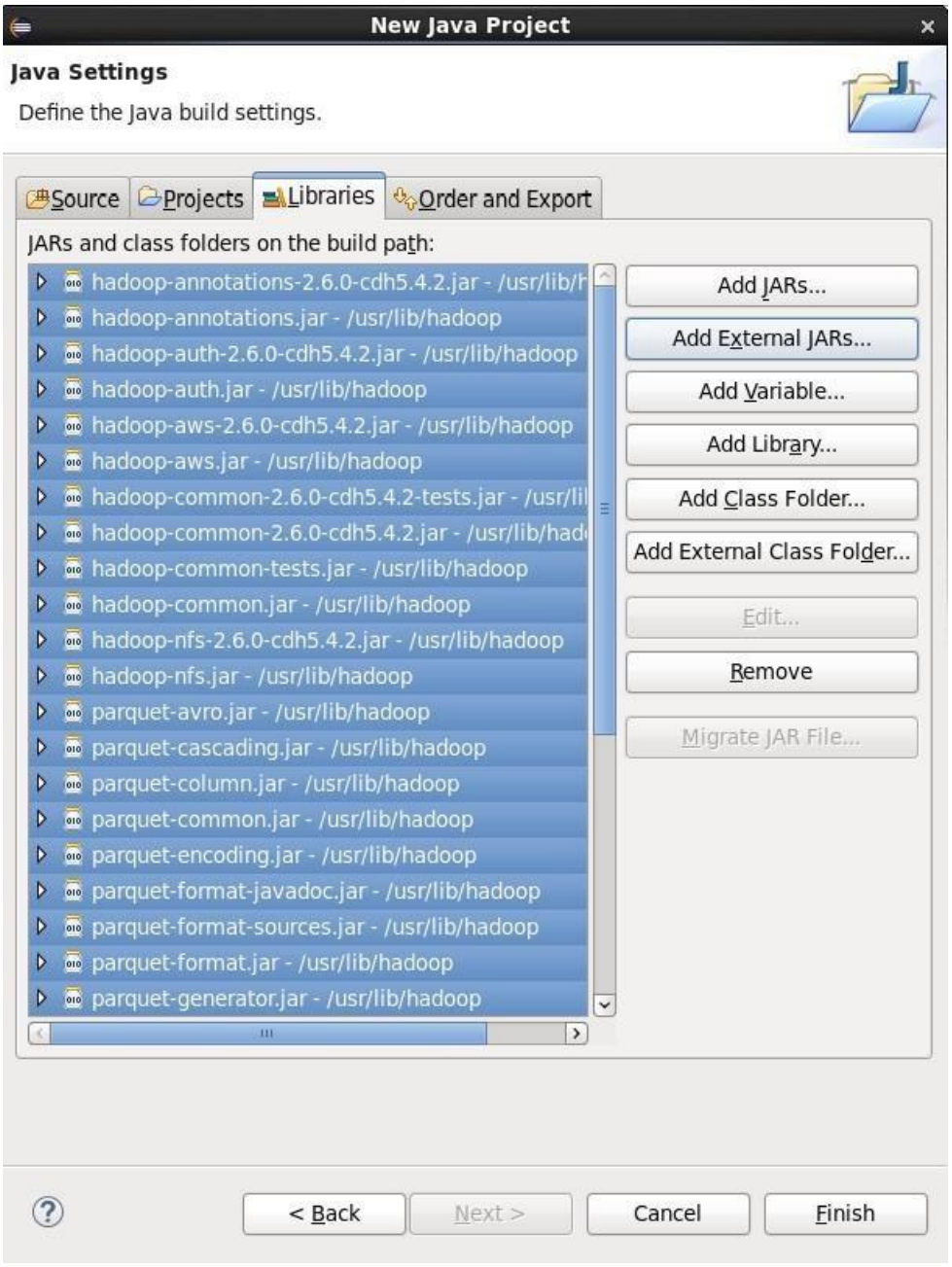

JAR Selection
usr
Places
Search
Recently Used
cloudera
Desktop
File System
Documents
Music
Pictures
Videos
Downloads
Name
Size
Modified
bin 06/09/2015
etc 09/23/2011
games 09/23/2011
include 06/09/2015
jars 06/09/2015
java 06/09/2015
lib 06/09/2015
lib64 06/09/2015
libexec 06/09/2015
local 06/09/2015
sbin 06/09/2015
share 06/09/2015
src 06/09/2015
tmp 06:13
Add
Remove
.jar;.zip
Cancel
OK

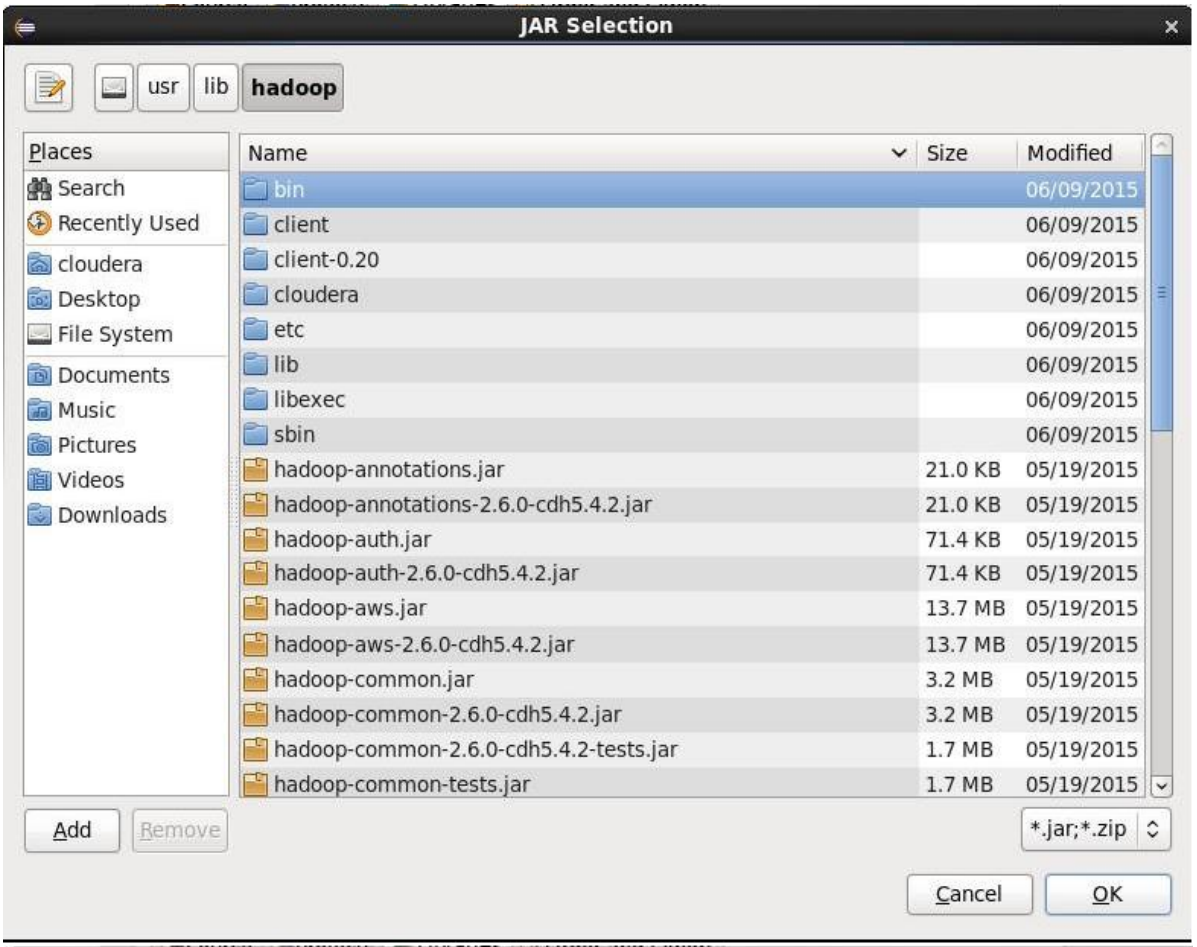
BIGDATA

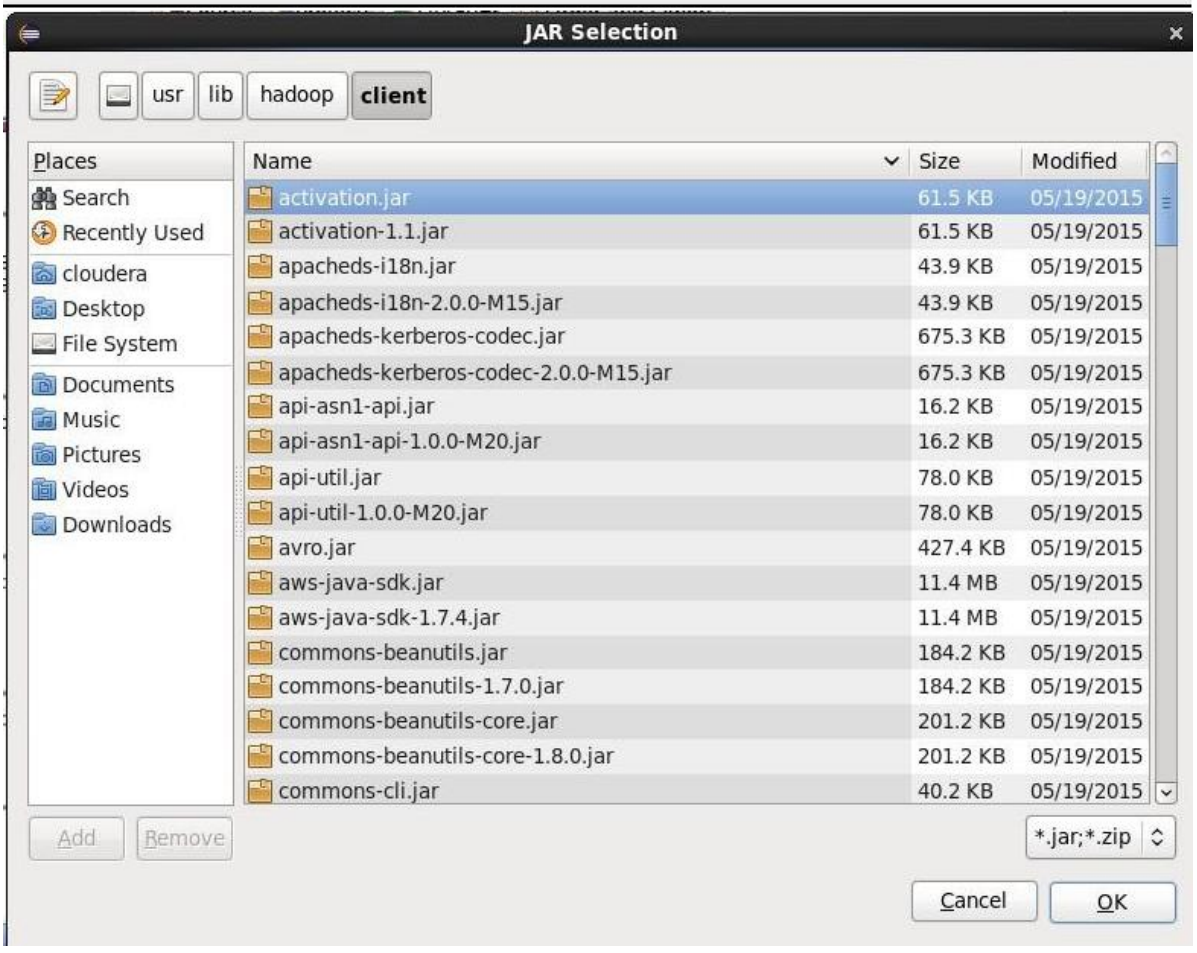

JAR Selection
usr
lib
Places
Search
Recently Used
cloudera
Desktop
File System
Documents
Music
Pictures
Videos
Downloads
Name
Size
Modified
anaconda-runtime 06/09/2015
avro 06/09/2015
bigtop-tomcat 06/09/2015
bigtop-utils 06/09/2015
bonobo 06/09/2015
ConsoleKit 06/09/2015
crunch 06/09/2015
cups 06/09/2015
dkms 06/09/2015
flume-ng 06/09/2015
games 09/23/2011
gcc 02/21/2013
hadoop 06/09/2015
hadoop-0.20-mapreduce 06/09/2015
hadoop-hdfs 06/09/2015
hadoop-httpfs 06/09/2015
hadoop-kms 06/09/2015
hadoop-mapreduce 06/09/2015
Add
Remove
.jar;.zip
Cancel
OK


JAR Selection
usr
lib
hadoop
Places
Search
Recently Used
cloudera
Desktop
File System
Documents
Music
Pictures
Videos
Downloads
Name
Size
Modified
bin 06/09/2015
client 06/09/2015
client-0.20 06/09/2015
cloudera 06/09/2015
etc 06/09/2015
lib 06/09/2015
libexec 06/09/2015
sbin 06/09/2015
hadoop-annotations.jar 21.0 KB 05/19/2015
hadoop-annotations-2.6.0-cdh5.4.2.jar 21.0 KB 05/19/2015
hadoop-auth.jar 71.4 KB 05/19/2015
hadoop-auth-2.6.0-cdh5.4.2.jar 71.4 KB 05/19/2015
hadoop-aws.jar 13.7 MB 05/19/2015
hadoop-aws-2.6.0-cdh5.4.2.jar 13.7 MB 05/19/2015
hadoop-common.jar 3.2 MB 05/19/2015
hadoop-common-2.6.0-cdh5.4.2.jar 3.2 MB 05/19/2015
hadoop-common-2.6.0-cdh5.4.2-tests.jar 1.7 MB 05/19/2015
hadoop-common-tests.jar 1.7 MB 05/19/2015
Add
Remove
.jar;.zip
Cancel
OK

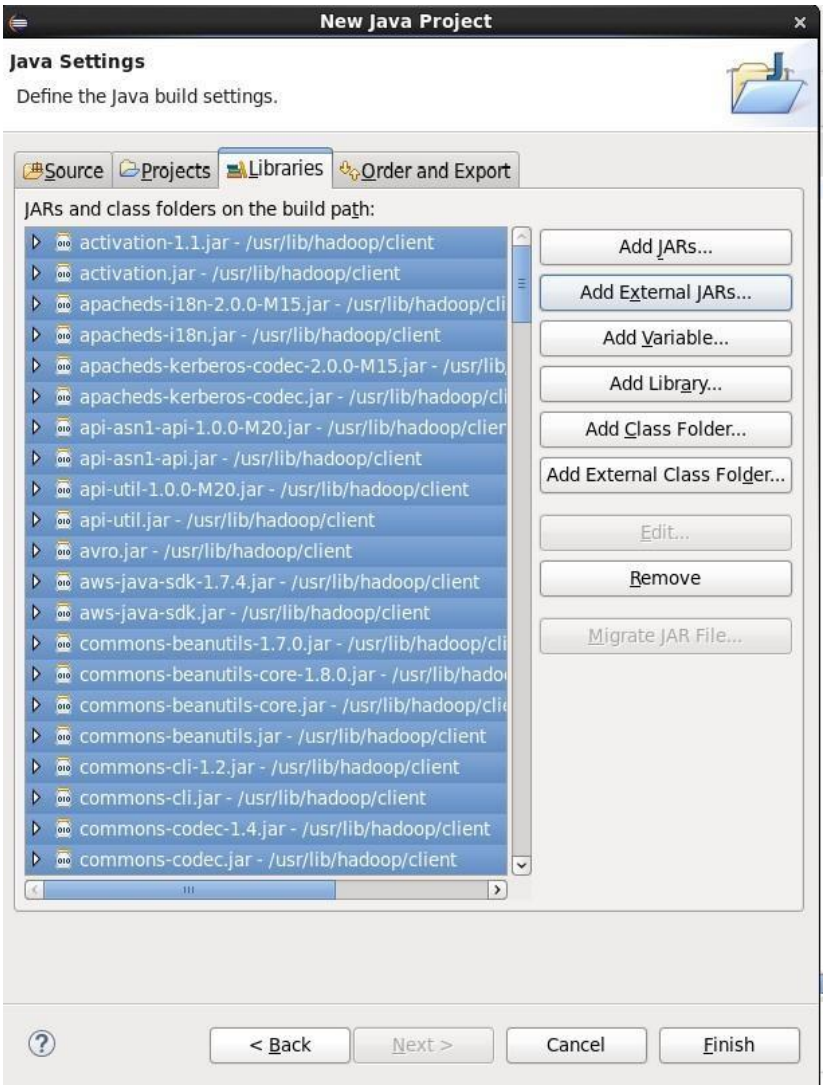
BIGDATA


New Java Project
Java Settings
Define the Java build settings.
Source
Projects
Libraries
Order and Export
JARs and class folders on the build path:
hadoop-annotations-2.6.0-cdh5.4.2.jar - /usr/lib/h
hadoop-annotations.jar - /usr/lib/hadoop
hadoop-auth-2.6.0-cdh5.4.2.jar - /usr/lib/hadoop
hadoop-auth.jar - /usr/lib/hadoop
hadoop-aws-2.6.0-cdh5.4.2.jar - /usr/lib/hadoop
hadoop-aws.jar - /usr/lib/hadoop
hadoop-common-2.6.0-cdh5.4.2-tests.jar - /usr/li
hadoop-common-2.6.0-cdh5.4.2.jar - /usr/lib/had
hadoop-common-tests.jar - /usr/lib/hadoop
hadoop-common.jar - /usr/lib/hadoop
hadoop-nfs-2.6.0-cdh5.4.2.jar - /usr/lib/hadoop
hadoop-nfs.jar - /usr/lib/hadoop
parquet-avro.jar - /usr/lib/hadoop
parquet-cascading.jar - /usr/lib/hadoop
parquet-column.jar - /usr/lib/hadoop
parquet-common.jar - /usr/lib/hadoop
parquet-encoding.jar - /usr/lib/hadoop
parquet-format-javadoc.jar - /usr/lib/hadoop
parquet-format-sources.jar - /usr/lib/hadoop
parquet-format.jar - /usr/lib/hadoop
parquet-generator.jar - /usr/lib/hadoop
Add JARs...
Add External JARs...
Add Variable...
Add Library...
Add Class Folder...
Add External Class Folder...
Edit...
Remove
Migrate JAR File...
< Back
Next >
Cancel
Finish


JAR Selection
usr
lib
hadoop
Places
Search
Recently Used
cloudera
Desktop
File System
Documents
Music
Pictures
Videos
Downloads
Name
Size
Modified
bin 06/09/2015
client 06/09/2015
client-0.20 06/09/2015
cloudera 06/09/2015
etc 06/09/2015
lib 06/09/2015
libexec 06/09/2015
sbin 06/09/2015
hadoop-annotations.jar 21.0 KB 05/19/2015
hadoop-annotations-2.6.0-cdh5.4.2.jar 21.0 KB 05/19/2015
hadoop-auth.jar 71.4 KB 05/19/2015
hadoop-auth-2.6.0-cdh5.4.2.jar 71.4 KB 05/19/2015
hadoop-aws.jar 13.7 MB 05/19/2015
hadoop-aws-2.6.0-cdh5.4.2.jar 13.7 MB 05/19/2015
hadoop-common.jar 3.2 MB 05/19/2015
hadoop-common-2.6.0-cdh5.4.2.jar 3.2 MB 05/19/2015
hadoop-common-2.6.0-cdh5.4.2-tests.jar 1.7 MB 05/19/2015
hadoop-common-tests.jar 1.7 MB 05/19/2015
Add
Remove
.jar;.zip
Cancel
OK


JAR Selection
usr
lib
hadoop
client
Places
Search
Recently Used
cloudera
Desktop
File System
Documents
Music
Pictures
Videos
Downloads
Name
Size
Modified
activation.jar 61.5 KB 05/19/2015
activation-1.1.jar 61.5 KB 05/19/2015
apacheds-i18n.jar 43.9 KB 05/19/2015
apacheds-i18n-2.0.0-M15.jar 43.9 KB 05/19/2015
apacheds-kerberos-codec.jar 675.3 KB 05/19/2015
apacheds-kerberos-codec-2.0.0-M15.jar 675.3 KB 05/19/2015
api-asn1-api.jar 16.2 KB 05/19/2015
api-asn1-api-1.0.0-M20.jar 16.2 KB 05/19/2015
api-util.jar 78.0 KB 05/19/2015
api-util-1.0.0-M20.jar 78.0 KB 05/19/2015
avro.jar 427.4 KB 05/19/2015
aws-java-sdk.jar 11.4 MB 05/19/2015
aws-java-sdk-1.7.4.jar 11.4 MB 05/19/2015
commons-beanutils.jar 184.2 KB 05/19/2015
commons-beanutils-1.7.0.jar 184.2 KB 05/19/2015
commons-beanutils-core.jar 201.2 KB 05/19/2015
commons-beanutils-core-1.8.0.jar 201.2 KB 05/19/2015
commons-cli.jar 40.2 KB 05/19/2015
Add
Remove
.jar;.zip
Cancel
OK

BIGDATA

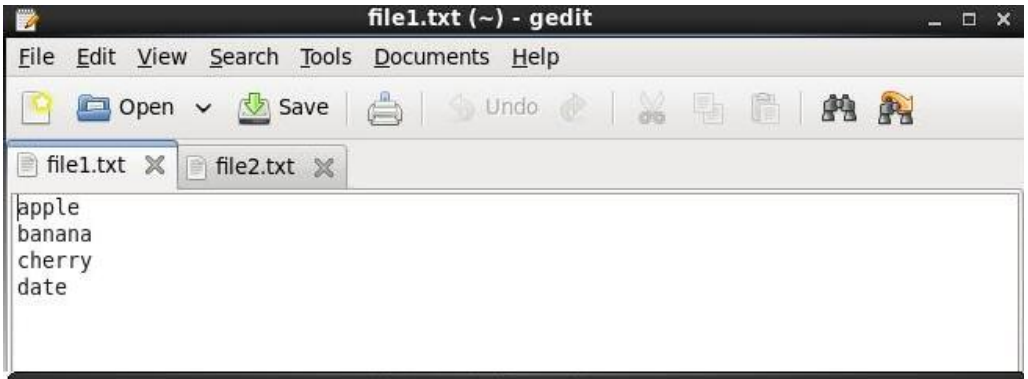

New Java Project
Java Settings
Define the Java build settings.
Source
Projects
Libraries
Order and Export
JARs and class folders on the build path:
activation-1.1.jar - /usr/lib/hadoop/client
activation.jar - /usr/lib/hadoop/client
apacheds-i18n-2.0.0-M15.jar - /usr/lib/hadoop/cli
apacheds-i18n.jar - /usr/lib/hadoop/client
apacheds-kerberos-codec-2.0.0-M15.jar - /usr/lib/
apacheds-kerberos-codec.jar - /usr/lib/hadoop/cli
api-asn1-api-1.0.0-M20.jar - /usr/lib/hadoop/clier
api-asn1-api.jar - /usr/lib/hadoop/client
api-util-1.0.0-M20.jar - /usr/lib/hadoop/client
api-util.jar - /usr/lib/hadoop/client
avro.jar - /usr/lib/hadoop/client
aws-java-sdk-1.7.4.jar - /usr/lib/hadoop/client
aws-java-sdk.jar - /usr/lib/hadoop/client
commons-beanutils-1.7.0.jar - /usr/lib/hadoop/cli
commons-beanutils-core-1.8.0.jar - /usr/lib/hado
commons-beanutils-core.jar - /usr/lib/hadoop/cli
commons-beanutils.jar - /usr/lib/hadoop/client
commons-cli-1.2.jar - /usr/lib/hadoop/client
commons-cli.jar - /usr/lib/hadoop/client
commons-codec-1.4.jar - /usr/lib/hadoop/client
commons-codec.jar - /usr/lib/hadoop/client
Add JARs...
Add External JARs...
Add Variable...
Add Library...
Add Class Folder...
Add External Class Folder...
Edit...
Remove
Migrate JAR File...
< Back
Next >
Cancel
Finish

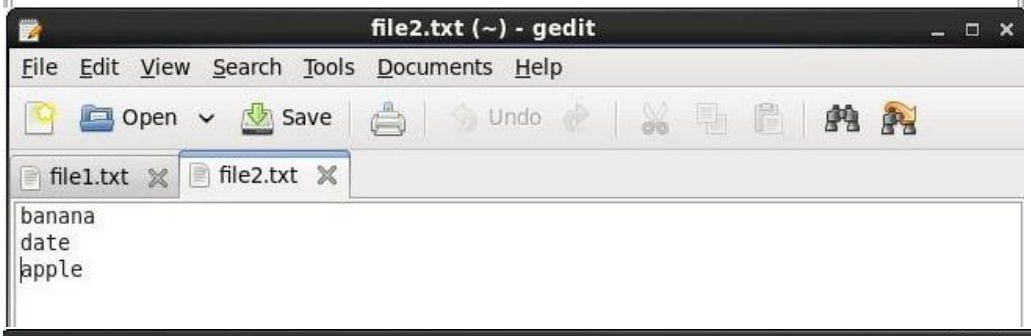
BIGDATA

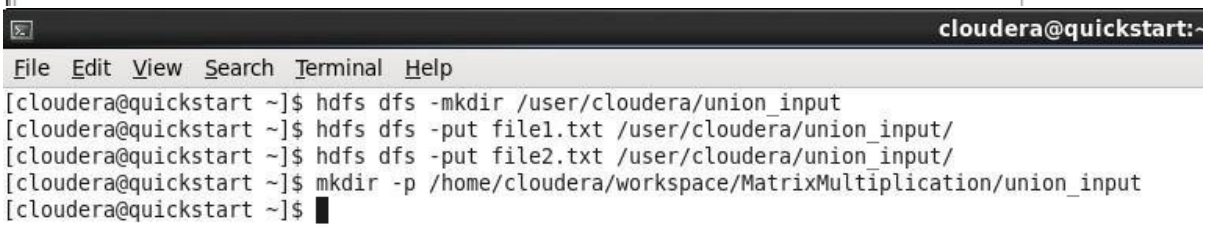
StubDriver.java | Union.java

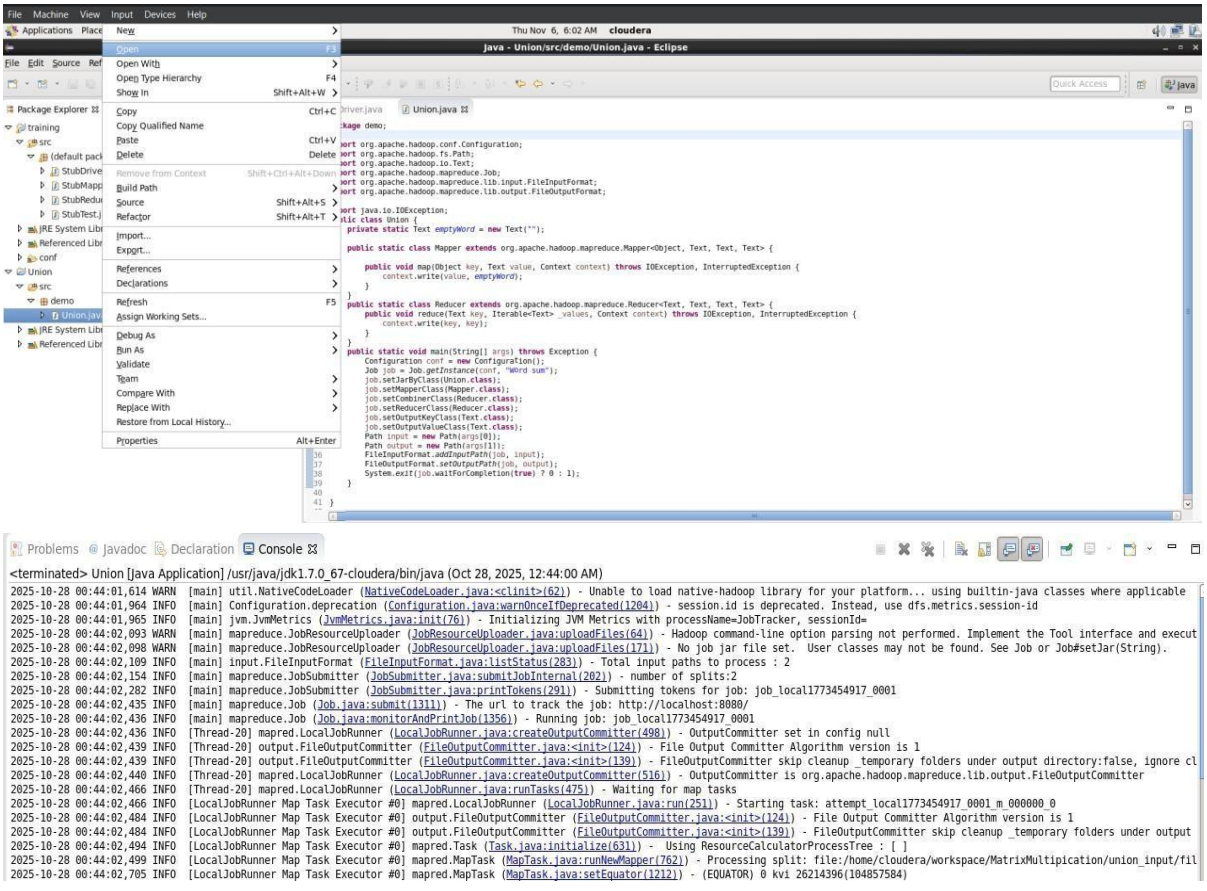
```java
package demo;

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

import java.io.IOException;
public class Union {
    private static Text emptyWord = new Text("");

    public static class Mapper extends org.apache.hadoop.mapreduce.Mapper<Object, Text, Text, Text> {

        public void map(Object key, Text value, Context context) throws IOException, InterruptedException {
            context.write(value, emptyWord);
        }
    }
    public static class Reducer extends org.apache.hadoop.mapreduce.Reducer<Text, Text, Text, Text> {
        public void reduce(Text key, Iterable<Text> _values, Context context) throws IOException, InterruptedException {
            context.write(key, key);
        }
    }
    public static void main(String[] args) throws Exception {
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "Word sum");
        job.setJarByClass(Union.class);
        job.setMapperClass(Mapper.class);
        job.setCombinerClass(Reducer.class);
        job.setReducerClass(Reducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(Text.class);
        Path input = new Path(args[0]);
        Path output = new Path(args[1]);
        FileInputFormat.addInputPath(job, input);
        FileOutputFormat.setOutputPath(job, output);
        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }

}
```

BIGDATA


file1.txt (~) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
file1.txt file2.txt
apple
banana
cherry
date

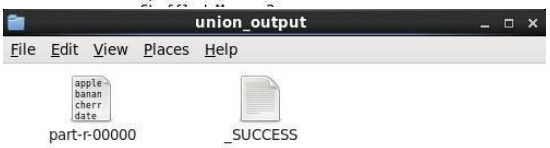

file2.txt (~) - gedit
File Edit View Search Tools Documents Help
Open Save Undo
file1.txt file2.txt
banana
date
apple

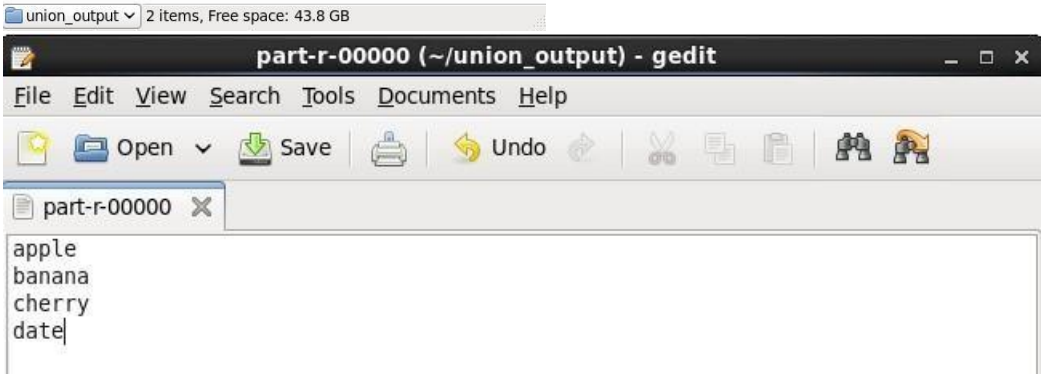

cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ hdfs dfs -mkdir /user/cloudera/union_input
[cloudera@quickstart ~]$ hdfs dfs -put file1.txt /user/cloudera/union_input/
[cloudera@quickstart ~]$ hdfs dfs -put file2.txt /user/cloudera/union_input/
[cloudera@quickstart ~]$ mkdir -p /home/cloudera/workspace/MatrixMultiplication/union_input
[cloudera@quickstart ~]$


java - Union/src/demo/Union.java - Eclipse
Problems Javadoc Declaration Console
<terminated> Union [Java Application] /usr/java/jdk1.7.0_67-cloudera/bin/java (Oct 28, 2025, 12:44:00 AM)
2025-10-28 00:44:01,614 WARN [main] util.NativeCodeLoader (NativeCodeLoader.java:<clinit>(62)) - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2025-10-28 00:44:01,964 INFO [main] Configuration.deprecation (Configuration.java:warnOnceIfDeprecated(1204)) - session.id is deprecated. Instead, use dfs.metrics.session-id
2025-10-28 00:44:01,965 INFO [main] jvm.JvmMetrics (JvmMetrics.java:init(76)) - Initializing JVM Metrics with processName=JobTracker, sessionId=
2025-10-28 00:44:02,093 WARN [main] mapreduce.JobResourceUploader (JobResourceUploader.java:uploadFiles(64)) - Hadoop command-line option parsing not performed. Implement the Tool interface and execut
2025-10-28 00:44:02,098 WARN [main] mapreduce.JobResourceUploader (JobResourceUploader.java:uploadFiles(171)) - No job jar file set. User classes may not be found. See Job or Job#setJar(String).
2025-10-28 00:44:02,109 INFO [main] input.FileInputFormat (FileInputFormat.java:listStatus(283)) - Total input paths to process : 2
2025-10-28 00:44:02,154 INFO [main] mapreduce.JobSubmitter (JobSubmitter.java:submitJobInternal(202)) - number of splits:2
2025-10-28 00:44:02,282 INFO [main] mapreduce.JobSubmitter (JobSubmitter.java:printTokens(291)) - Submitting tokens for job: job_local1773454917_0001
2025-10-28 00:44:02,435 INFO [main] mapreduce.Job (Job.java:submit(1311)) - The url to track the job: http://localhost:8080/
2025-10-28 00:44:02,436 INFO [main] mapreduce.Job (Job.java:monitorAndPrintJob(1356)) - Running job: job_local1773454917_0001
2025-10-28 00:44:02,436 INFO [Thread-20] mapred.LocalJobRunner (LocalJobRunner.java:createOutputCommitter(498)) - OutputCommitter set in config null
2025-10-28 00:44:02,439 INFO [Thread-20] output.FileOutputCommitter (FileOutputCommitter.java:<init>(124)) - File Output Committer Algorithm version is 1
2025-10-28 00:44:02,439 INFO [Thread-20] output.FileOutputCommitter (FileOutputCommitter.java:<init>(139)) - FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cl
2025-10-28 00:44:02,440 INFO [Thread-20] mapred.LocalJobRunner (LocalJobRunner.java:createOutputCommitter(516)) - OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2025-10-28 00:44:02,466 INFO [Thread-20] mapred.LocalJobRunner (LocalJobRunner.java:runTasks(475)) - Waiting for map tasks
2025-10-28 00:44:02,466 INFO [LocalJobRunner Map Task Executor #0] mapred.LocalJobRunner (LocalJobRunner.java:run(251)) - Starting task: attempt_local1773454917_0001_m_000000_0
2025-10-28 00:44:02,484 INFO [LocalJobRunner Map Task Executor #0] output.FileOutputCommitter (FileOutputCommitter.java:<init>(124)) - File Output Committer Algorithm version is 1
2025-10-28 00:44:02,484 INFO [LocalJobRunner Map Task Executor #0] output.FileOutputCommitter (FileOutputCommitter.java:<init>(139)) - FileOutputCommitter skip cleanup _temporary folders under output
2025-10-28 00:44:02,494 INFO [LocalJobRunner Map Task Executor #0] mapred.Task (Task.java:initialize(631)) - Using ResourceCalculatorProcessTree : [ ]
2025-10-28 00:44:02,499 INFO [LocalJobRunner Map Task Executor #0] mapred.MapTask (MapTask.java:runNewMapper(762)) - Processing split: file:/home/cloudera/workspace/MatrixMultipication/union_input/fil
2025-10-28 00:44:02,705 INFO [LocalJobRunner Map Task Executor #0] mapred.MapTask (MapTask.java:setEquator(1212)) - (EQUATOR) 0 kvi 26214396(104857584)

BIGDATA

```
2025-10-28 00:44:03,441 INFO  [main] mapreduce.Job (Job.java:monitorAndPrintJob(1384)) -  map 100% reduce 100%
2025-10-28 00:44:03,442 INFO  [main] mapreduce.Job (Job.java:monitorAndPrintJob(1395)) - Job job_local1773454917_0001 completed successfully
2025-10-28 00:44:03,450 INFO  [main] mapreduce.Job (Job.java:monitorAndPrintJob(1402)) - Counters: 30
        File System Counters
                FILE: Number of bytes read=1892
                FILE: Number of bytes written=1000096
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
        Map-Reduce Framework
                Map input records=6
                Map output records=6
                Map output bytes=44
                Map output materialized bytes=68
                Input split bytes=272
                Combine input records=0
                Combine output records=0
                Reduce input groups=4
                Reduce shuffle bytes=68
                Reduce input records=6
                Reduce output records=4
                Spilled Records=12
```



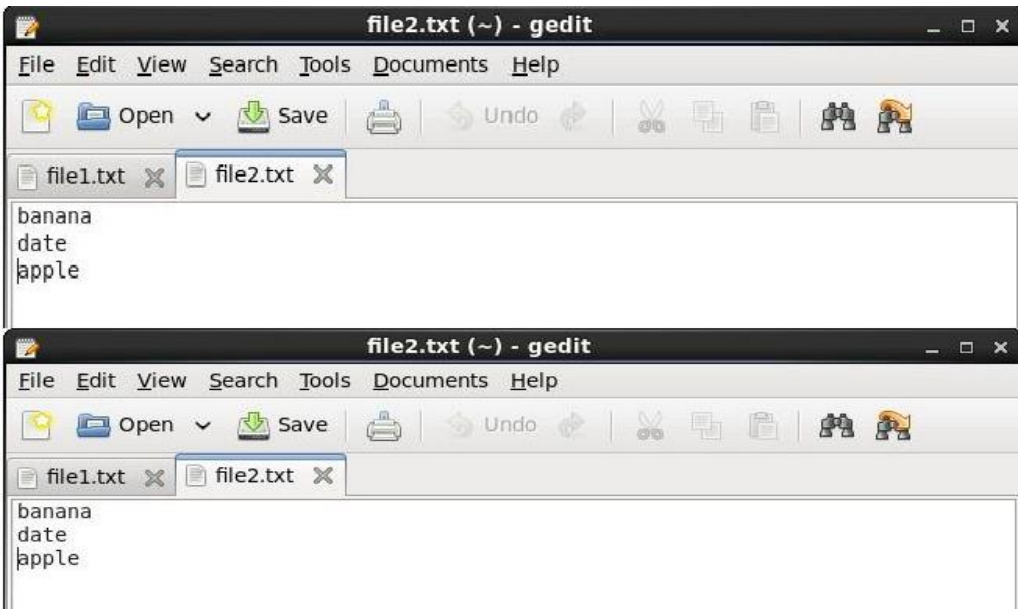


BIGDATA

Write a program in Map Reduce for Intersection operation

StubDriver.java | Union.java | Intersection.java

```
package demo;

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

public class Intersection {
    public static class Mapper extends org.apache.hadoop.mapreduce.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 {
            context.write(value, one);
        }
    }
    public static class Combiner extends org.apache.hadoop.mapreduce.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++;
            }
            result.set(sum);
            context.write(key, result);
        }
    }
    public static class Reducer extends org.apache.hadoop.mapreduce.Reducer<Text, IntWritable, Text, Text> {
        public void reduce(Text key, Iterable<IntWritable> values, Context context ) throws IOException, InterruptedException {
            int sum = 0;
            for (IntWritable val : values) {
                sum++;
            }
            if (sum > 1) {
                context.write(key, key);
            }
        }
    }
    public static void main(String[] args) throws
    Exception {
        Configuration conf = new Configuration();
        Job job = Job.getInstance(conf, "word count");
        job.setJarByClass(Intersection.class);
        job.setMapperClass(Mapper.class);
        job.setReducerClass(Reducer.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);
    }
}
```

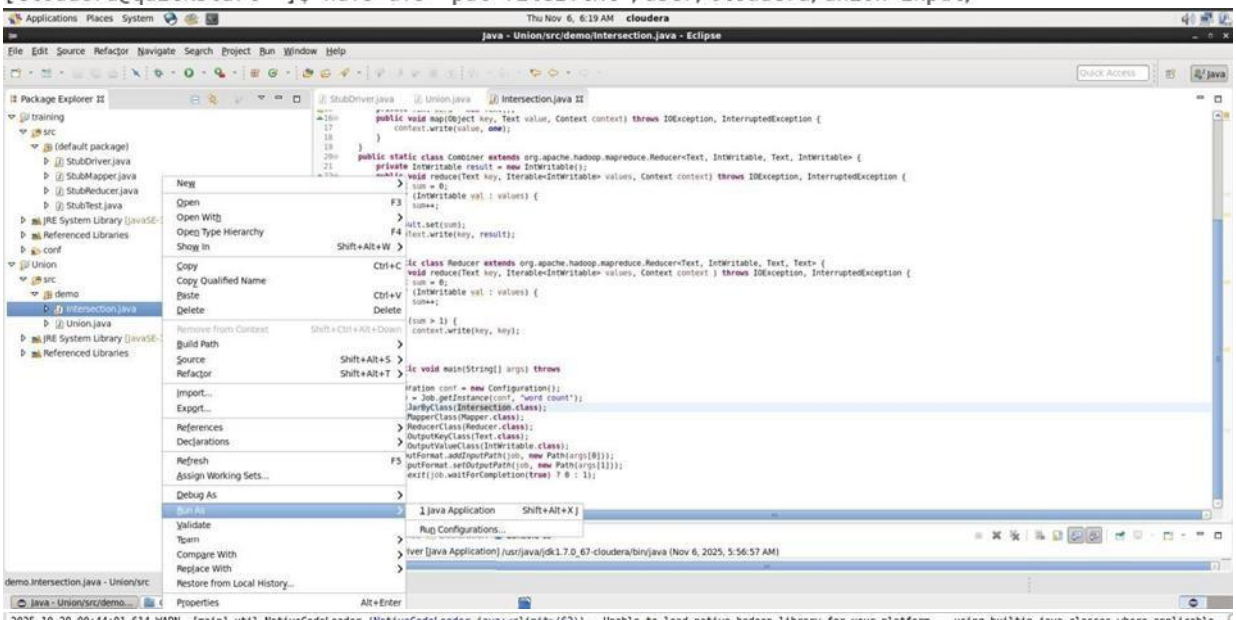


BIGDATA

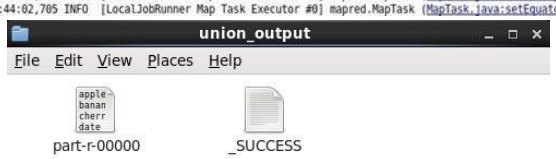
cloudera@quickstart:~

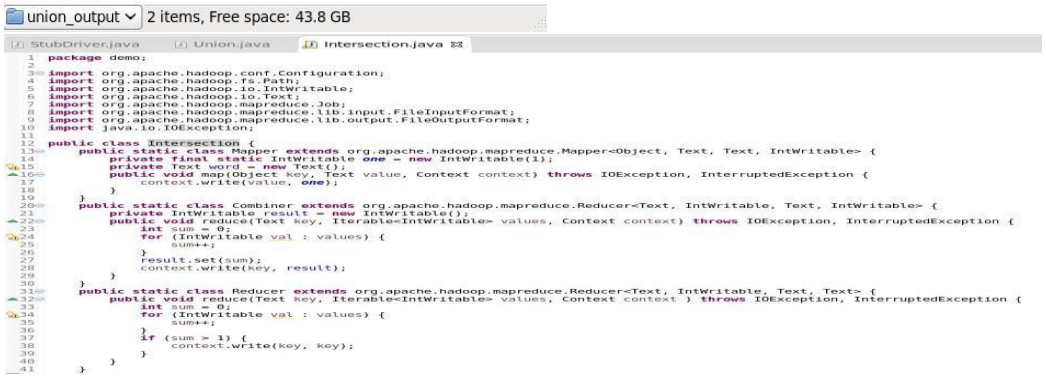
File Edit View Search Terminal Help

```
[cloudera@quickstart ~]$ hdfs dfs -mkdir /user/cloudera/union_input
[cloudera@quickstart ~]$ hdfs dfs -put file1.txt /user/cloudera/union_input/
[cloudera@quickstart ~]$ hdfs dfs -put file2.txt /user/cloudera/union_input/
```




```
2025-10-28 00:44:01,614 WARN  [main] util.NativeCodeLoader (NativeCodeLoader.java:<clinit>(62)) - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2025-10-28 00:44:01,964 INFO  [main] Configuration.deprecation (Configuration.java:warnOnceIfDeprecated(1204)) - session.id is deprecated. Instead, use dfs.metrics.session-id
2025-10-28 00:44:01,965 INFO  [main] jvm.JvmMetrics (JvmMetrics.java:init(76)) - Initializing JVM Metrics with processName=JobTracker, sessionId=
2025-10-28 00:44:02,093 WARN  [main] mapreduce.JobResourceUploader (JobResourceUploader.java:uploadFiles(64)) - Hadoop command-line option parsing not performed. Implement the Tool interface and execut
2025-10-28 00:44:02,098 WARN  [main] mapreduce.JobResourceUploader (JobResourceUploader.java:uploadFiles(171)) - No job jar file set.  User classes may not be found. See Job or Job#setJar(String).
2025-10-28 00:44:02,109 INFO  [main] input.FileInputFormat (FileInputFormat.java:listStatus(283)) - Total input paths to process : 2
2025-10-28 00:44:02,154 INFO  [main] mapreduce.JobSubmitter (JobSubmitter.java:submitJobInternal(202)) - number of splits:2
2025-10-28 00:44:02,282 INFO  [main] mapreduce.JobSubmitter (JobSubmitter.java:printTokens(291)) - Submitting tokens for job: job_local1773454917_0001
2025-10-28 00:44:02,435 INFO  [main] mapreduce.Job (Job.java:submit(1311)) - The url to track the job: http://localhost:8080/
2025-10-28 00:44:02,436 INFO  [main] mapreduce.Job (Job.java:monitorAndPrintJob(1356)) - Running job: job_local1773454917_0001
2025-10-28 00:44:02,436 INFO  [Thread-20] mapred.LocalJobRunner (LocalJobRunner.java:createOutputCommitter(498)) - OutputCommitter set in config null
2025-10-28 00:44:02,439 INFO  [Thread-20] output.FileOutputCommitter (FileOutputCommitter.java:<init>(124)) - File Output Committer Algorithm version is 1
2025-10-28 00:44:02,439 INFO  [Thread-20] output.FileOutputCommitter (FileOutputCommitter.java:<init>(139)) - FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cl
2025-10-28 00:44:02,440 INFO  [Thread-20] mapred.LocalJobRunner (LocalJobRunner.java:createOutputCommitter(516)) - OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2025-10-28 00:44:02,466 INFO  [Thread-20] mapred.LocalJobRunner (LocalJobRunner.java:runTasks(475)) - Waiting for map tasks
2025-10-28 00:44:02,466 INFO  [LocalJobRunner Map Task Executor #0] mapred.LocalJobRunner (LocalJobRunner.java:run(251)) - Starting task: attempt_local1773454917_0001_m_000000_0
2025-10-28 00:44:02,484 INFO  [LocalJobRunner Map Task Executor #0] output.FileOutputCommitter (FileOutputCommitter.java:<init>(124)) - File Output Committer Algorithm version is 1
2025-10-28 00:44:02,484 INFO  [LocalJobRunner Map Task Executor #0] output.FileOutputCommitter (FileOutputCommitter.java:<init>(139)) - FileOutputCommitter skip cleanup _temporary folders under output
2025-10-28 00:44:02,494 INFO  [LocalJobRunner Map Task Executor #0] mapred.Task (Task.java:initialize(631)) -  Using ResourceCalculatorProcessTree : [ ]
2025-10-28 00:44:02,499 INFO  [LocalJobRunner Map Task Executor #0] mapred.MapTask (MapTask.java:runNewMapper(762)) - Processing split: file:/home/cloudera/workspace/MatrixMultipication/union_input/fil
2025-10-28 00:44:02,705 INFO  [LocalJobRunner Map Task Executor #0] mapred.MapTask (MapTask.java:setEquator(1212)) - (EQUATOR) 0 kv1 26214396(104857584)
```

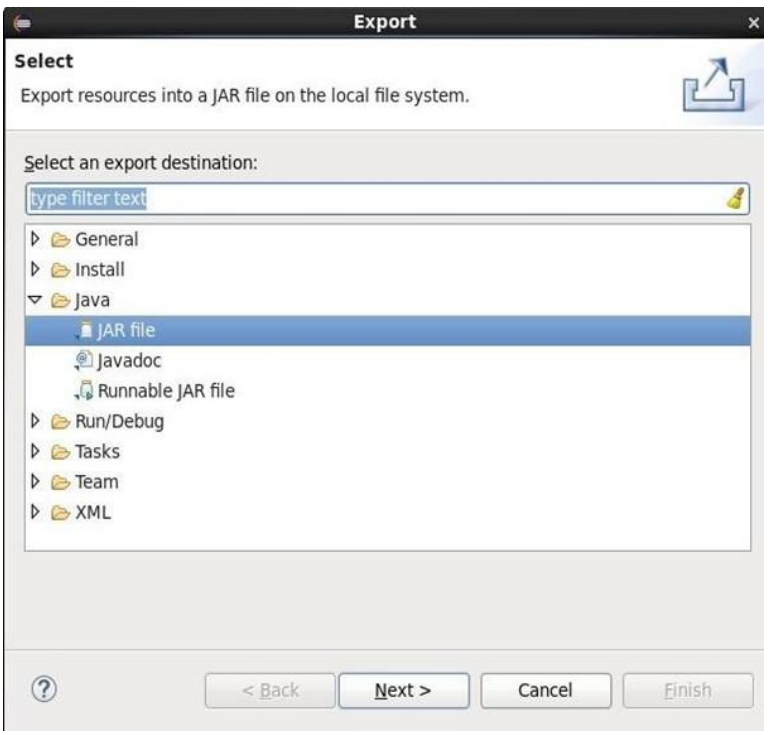


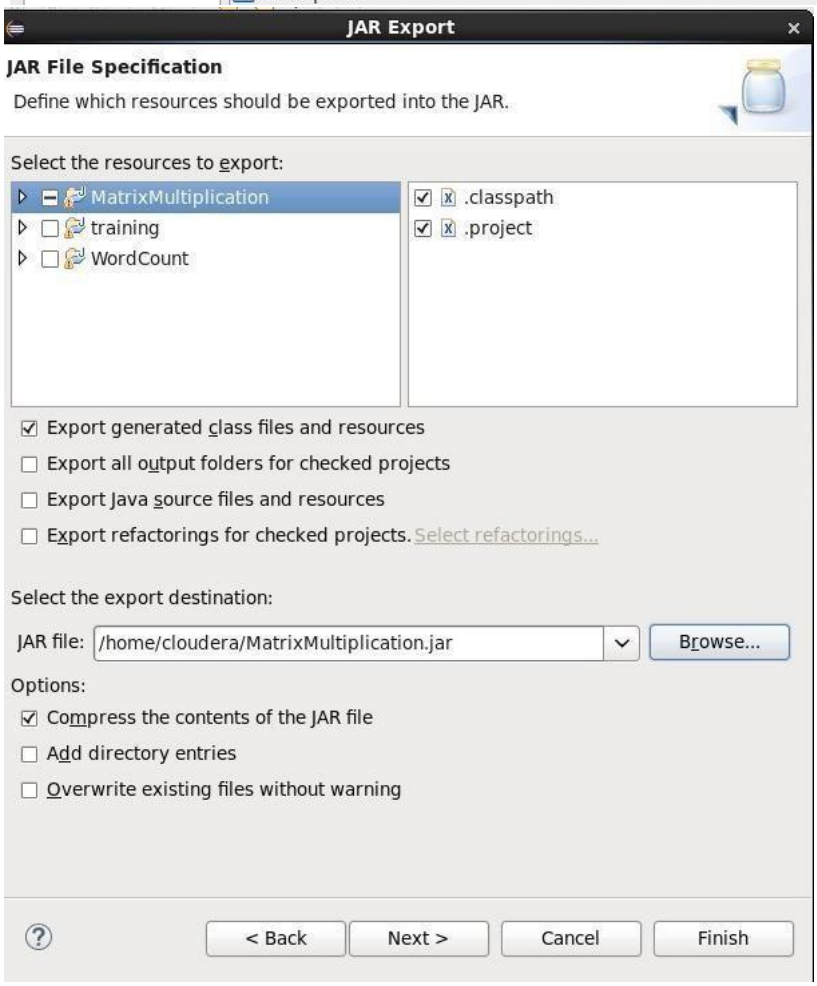



```
package demo;

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

public class Intersection {
    public static class Mapper extends org.apache.hadoop.mapreduce.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 {
            context.write(value, one);
        }
    }
    public static class Combiner extends org.apache.hadoop.mapreduce.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++;
            }
            result.set(sum);
            context.write(key, result);
        }
    }
    public static class Reducer extends org.apache.hadoop.mapreduce.Reducer<Text, IntWritable, Text, Text> {
        public void reduce(Text key, Iterable<IntWritable> values, Context context ) throws IOException, InterruptedException {
            int sum = 0;
            for (IntWritable val : values) {
                sum++;
            }
            if (sum > 1) {
                context.write(key, key);
            }
        }
    }
```

Write a program in Map Reduce for Matrix Multiplication.

Step 1: Open virtual box and then start cloudera quickstart
Step 2: Open eclipse present on the cloudera desktop
Step 3: Create java project
File->New-> Java Project
Give project name: MatrixMultiplication
Click Next
Step 4: Add Hadoop libraries to project
Select Libraries tab->click on Add Externa Jars
File System->user->lib->Hadoop
Select all library(jar) files->ok
Again, click on Add External Jars
File System->user->lib->Hadoop->client
Select all library(jar) files from client->ok->finish
Step 5: Write java code for matrix multiplication using mapreduce
Right click on src folder of project MatrixMultiplication
New->class
Write class name MatrixMultiplicationMapper
Click Finish
Write the code for MatrixMultiplicationMapper
Same way create classes MatrixMultiplicationReducer and MatrixMultiplicationDriver
Step 6: Export the project as jar
Right click on project MatrixMultiplication and select Export>> Java>>JAR file>>Next





Verify the jar file through command line open terminal give command ls

step 7: Move the jar file to the Hadoop file system hdfs dfs -put matrixmultiplication.jar /user/cloudera

hdfs dfs -ls

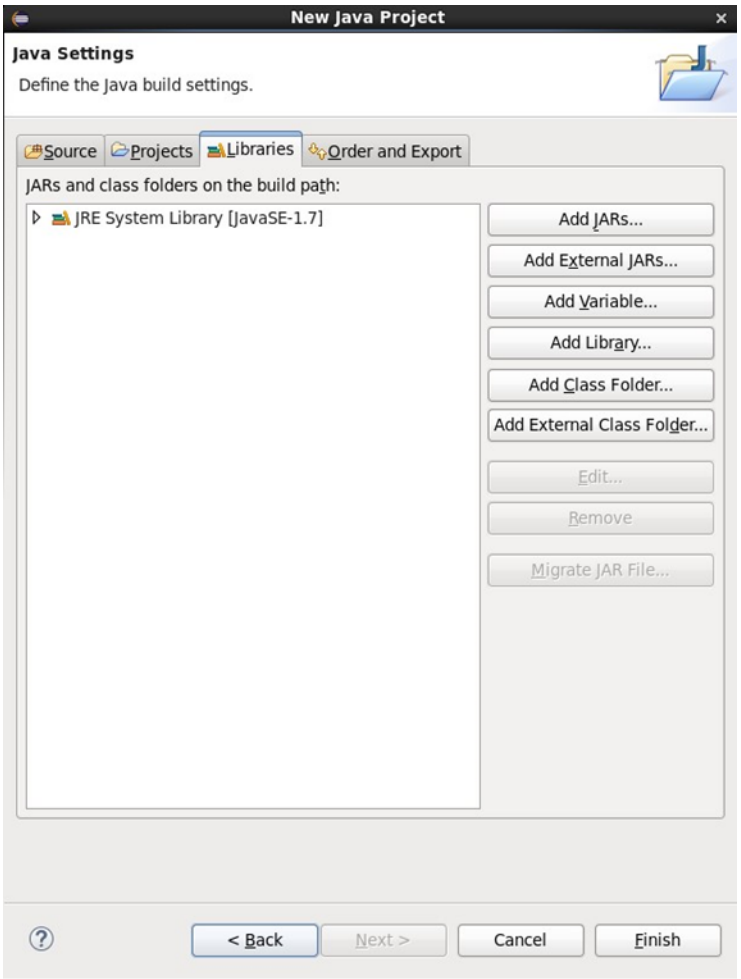
Step 8: Create the input file for the MapReduce program First input file

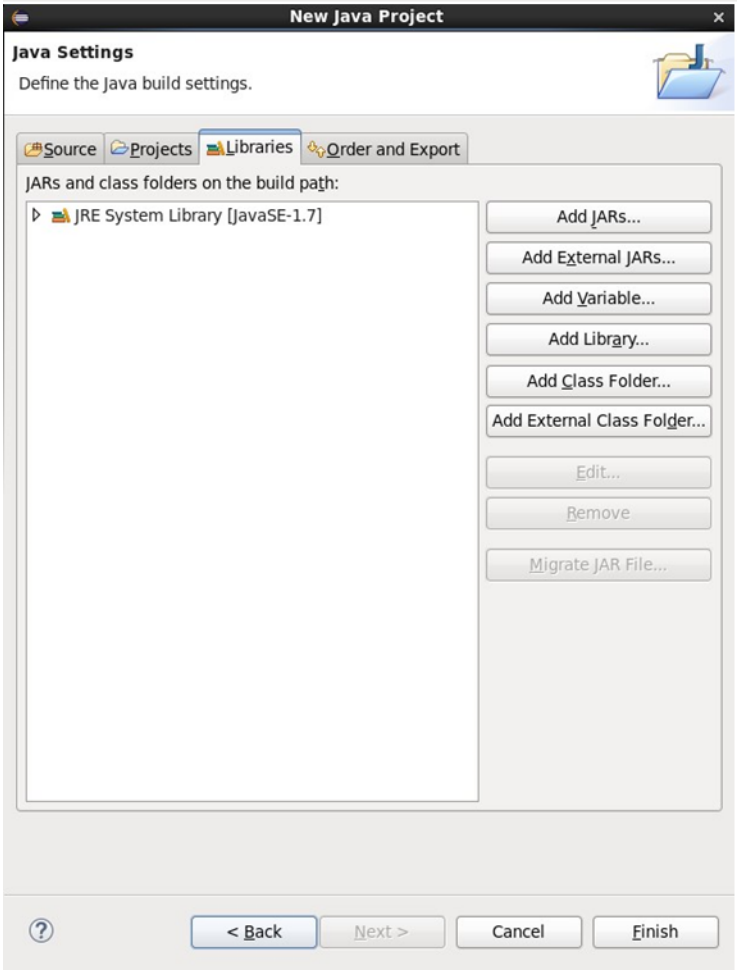
Command: cat > myMMatrix.txt

Enter data in input file
M,0,0,1
M,0,1,2 M,1,0,3 M,1,1,4 and press enter and ctrl z

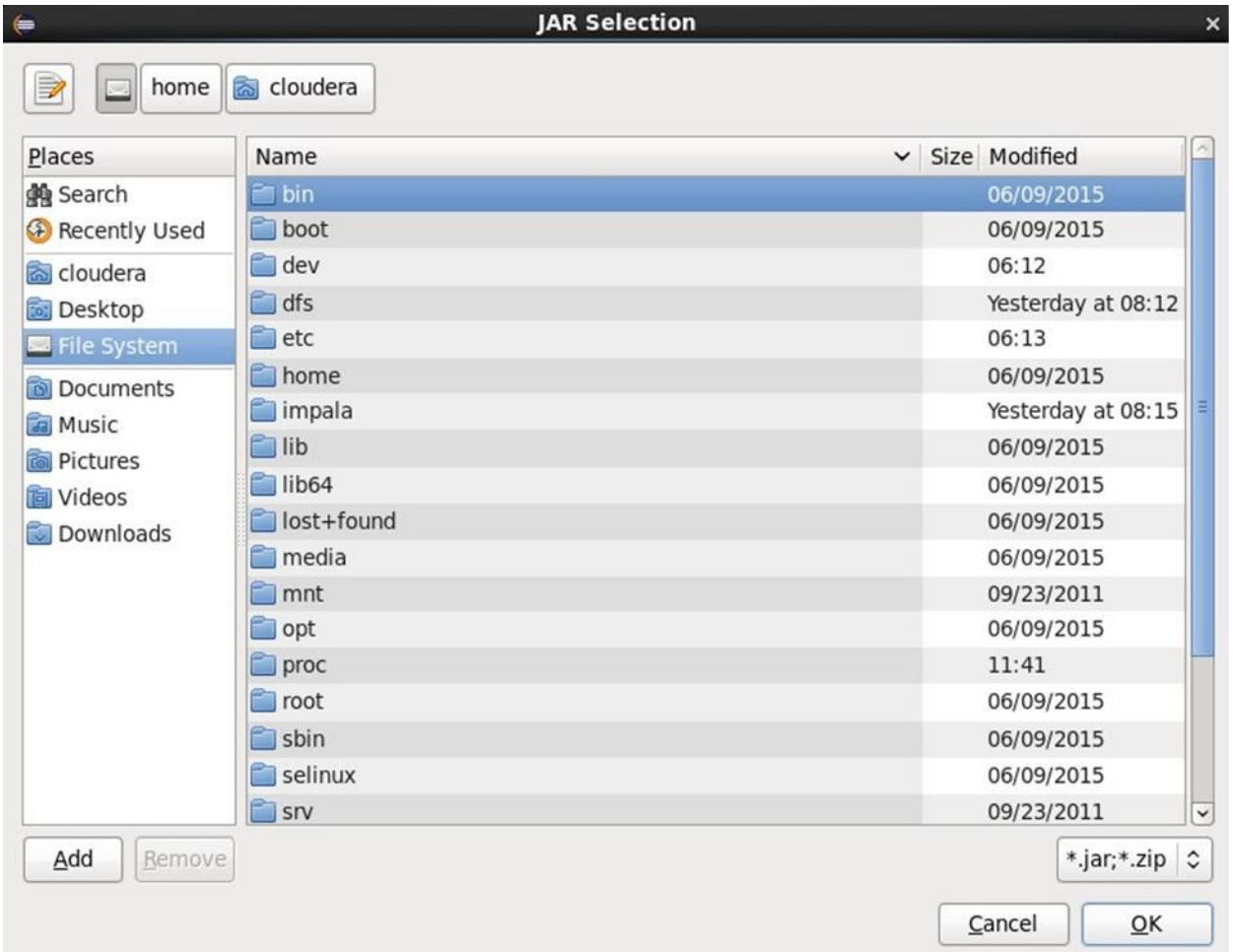
Command: cat myMMatrix.txt
Second input file
Command: cat > myNMatrix.txt
Enter data in input file
N,0,0,3
N,0,1,6
N,1,0,4 N,1,1,2 and press enter and ctrl z
Command: cat myNMatrix.txt
Step 9: Create the input directory and move input files into it. hdfs dfs –mkdir /user/cloudera/matrixInput hdfs dfs -put myMMatrix.txt /user/cloudera/matrixInput/ hdfs dfs -put myNMatrix.txt /user/cloudera/matrixInput/ hdfs dfs -ls
step 10: Run mapreduce program on Hadoop syntax: hadoop jar jarfilename.jar classname inputfoldername outputfoldername command: hadoop jar matrixmultiplication.jar MatrixMultiplicationDriver matrixInput matrixOutput
step 11: view output directory hdfs dfs –ls hdfs dfs -ls /user/cloudera/matrixOutput
step 12: view the output file hdfs dfs -cat /user/cloudera/matrixOutput/part-r-00000

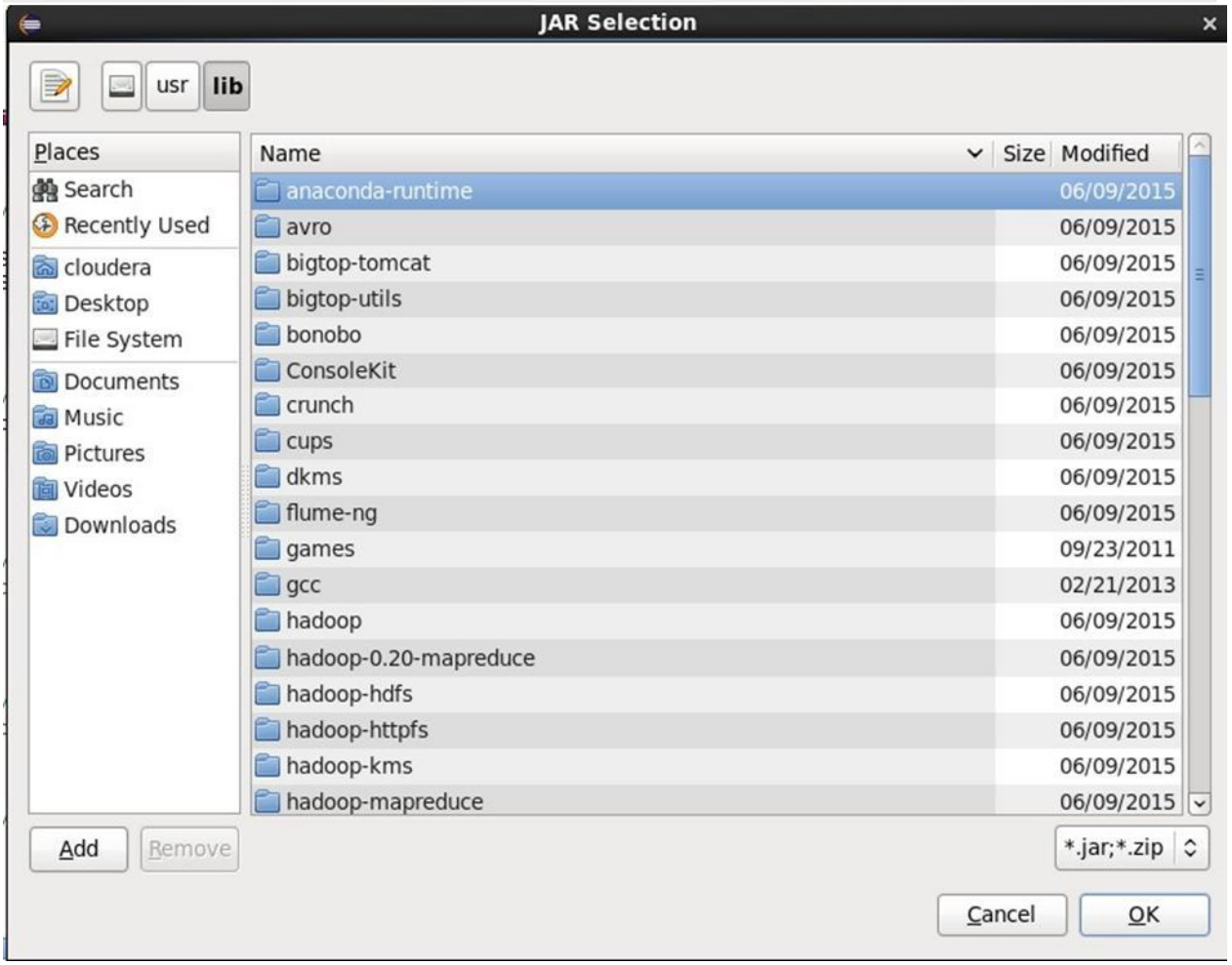
BIGDATA


New Java Project
Java Settings
Define the Java build settings.
Source
Projects
Libraries
Order and Export
JARs and class folders on the build path:
JRE System Library [JavaSE-1.7]
Add JARs...
Add External JARs...
Add Variable...
Add Library...
Add Class Folder...
Add External Class Folder...
Edit...
Remove
Migrate JAR File...
< Back
Next >
Cancel
Finish

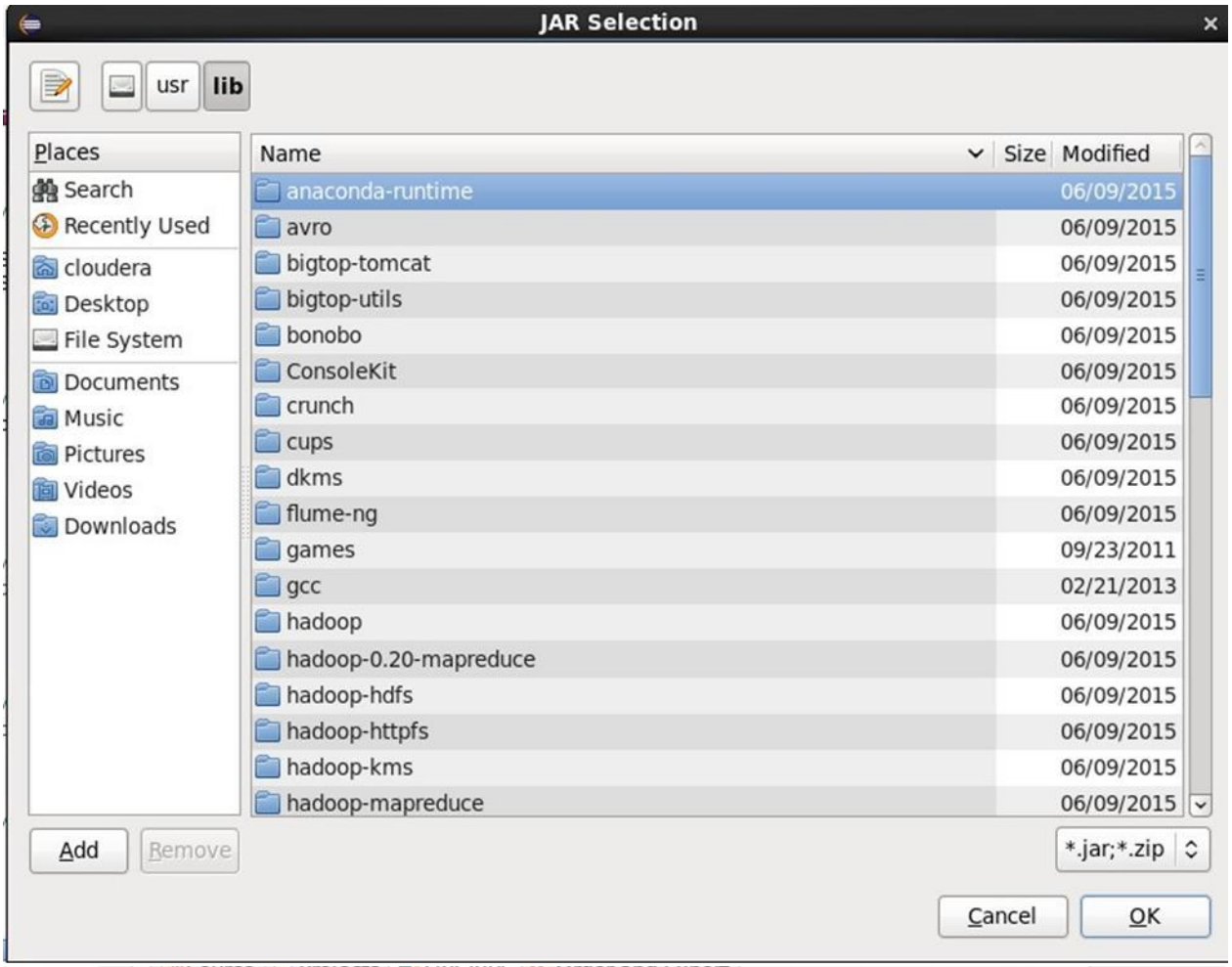

New Java Project
Java Settings
Define the Java build settings.
Source
Projects
Libraries
Order and Export
JARs and class folders on the build path:
JRE System Library [JavaSE-1.7]
Add JARs...
Add External JARs...
Add Variable...
Add Library...
Add Class Folder...
Add External Class Folder...
Edit...
Remove
Migrate JAR File...
< Back
Next >
Cancel
Finish

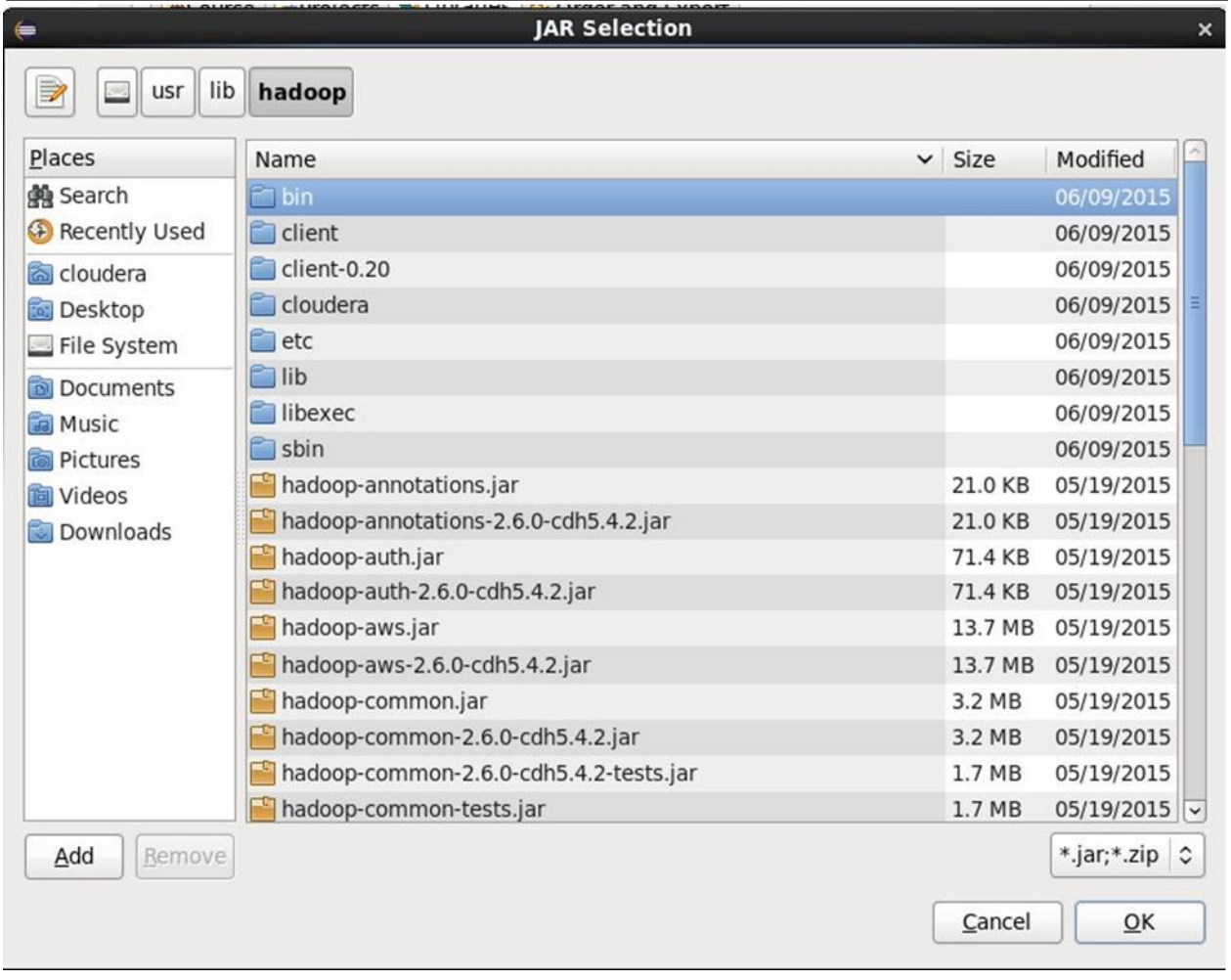
BIGDATA


JAR Selection
home
cloudera
Places
Search
Recently Used
cloudera
Desktop
File System
Documents
Music
Pictures
Videos
Downloads
Name
Size
Modified
bin 06/09/2015
boot 06/09/2015
dev 06:12
dfs Yesterday at 08:12
etc 06:13
home 06/09/2015
impala Yesterday at 08:15
lib 06/09/2015
lib64 06/09/2015
lost+found 06/09/2015
media 06/09/2015
mnt 09/23/2011
opt 06/09/2015
proc 11:41
root 06/09/2015
sbin 06/09/2015
selinux 06/09/2015
srv 09/23/2011
Add
Remove
.jar;.zip
Cancel
OK

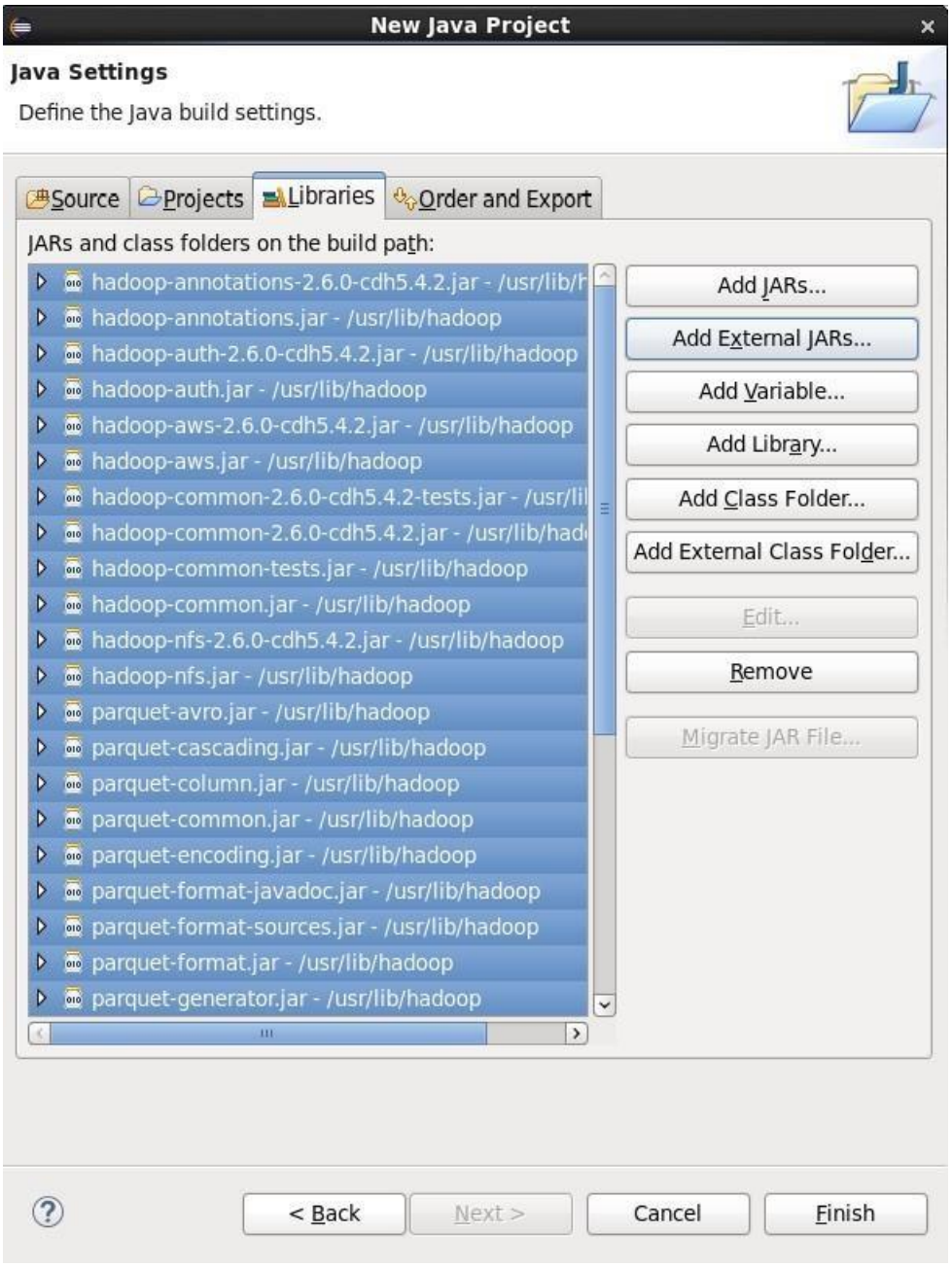

JAR Selection
usr
lib
Places
Search
Recently Used
cloudera
Desktop
File System
Documents
Music
Pictures
Videos
Downloads
Name
Size
Modified
anaconda-runtime 06/09/2015
avro 06/09/2015
bigtop-tomcat 06/09/2015
bigtop-utils 06/09/2015
bonobo 06/09/2015
ConsoleKit 06/09/2015
crunch 06/09/2015
cups 06/09/2015
dkms 06/09/2015
flume-ng 06/09/2015
games 09/23/2011
gcc 02/21/2013
hadoop 06/09/2015
hadoop-0.20-mapreduce 06/09/2015
hadoop-hdfs 06/09/2015
hadoop-httpfs 06/09/2015
hadoop-kms 06/09/2015
hadoop-mapreduce 06/09/2015
Add
Remove
.jar;.zip
Cancel
OK

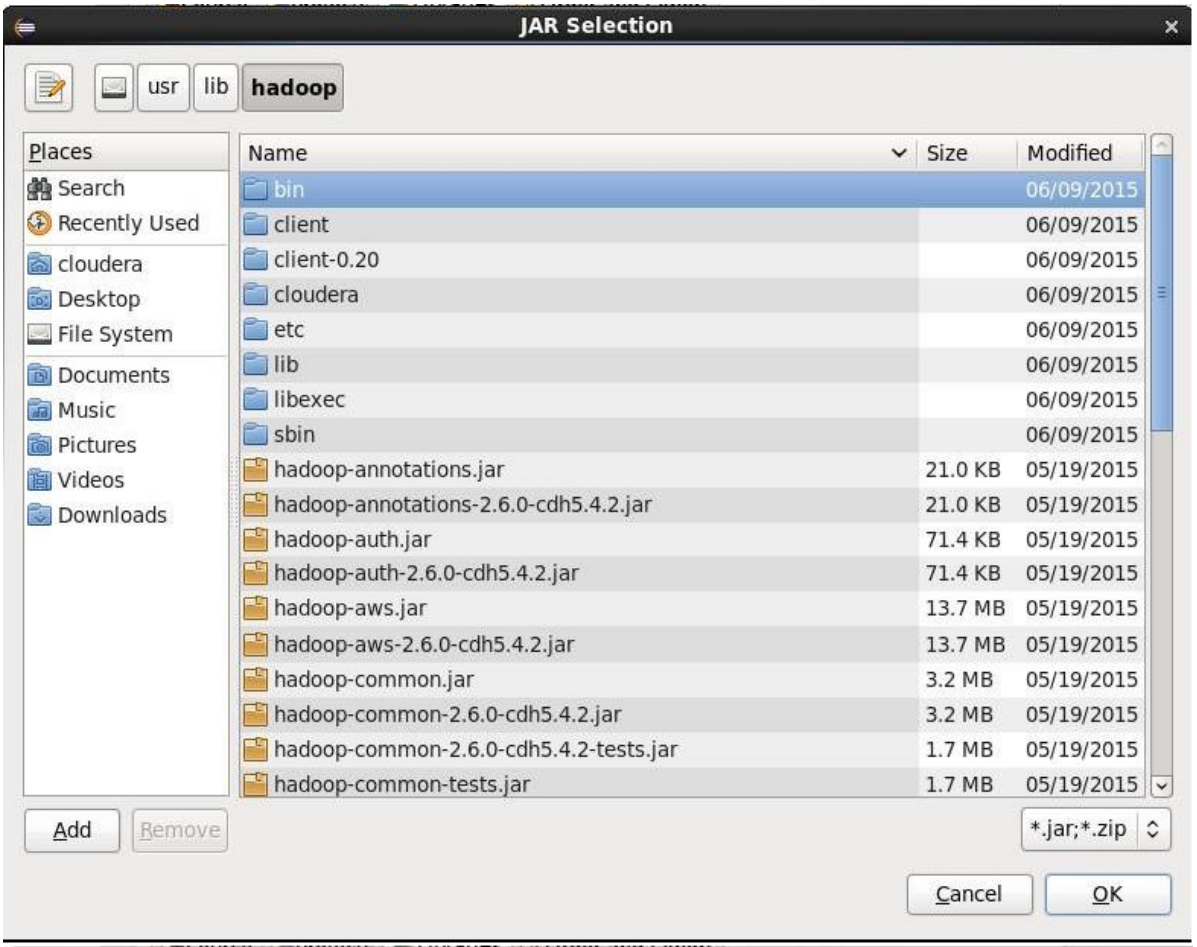
BIGDATA

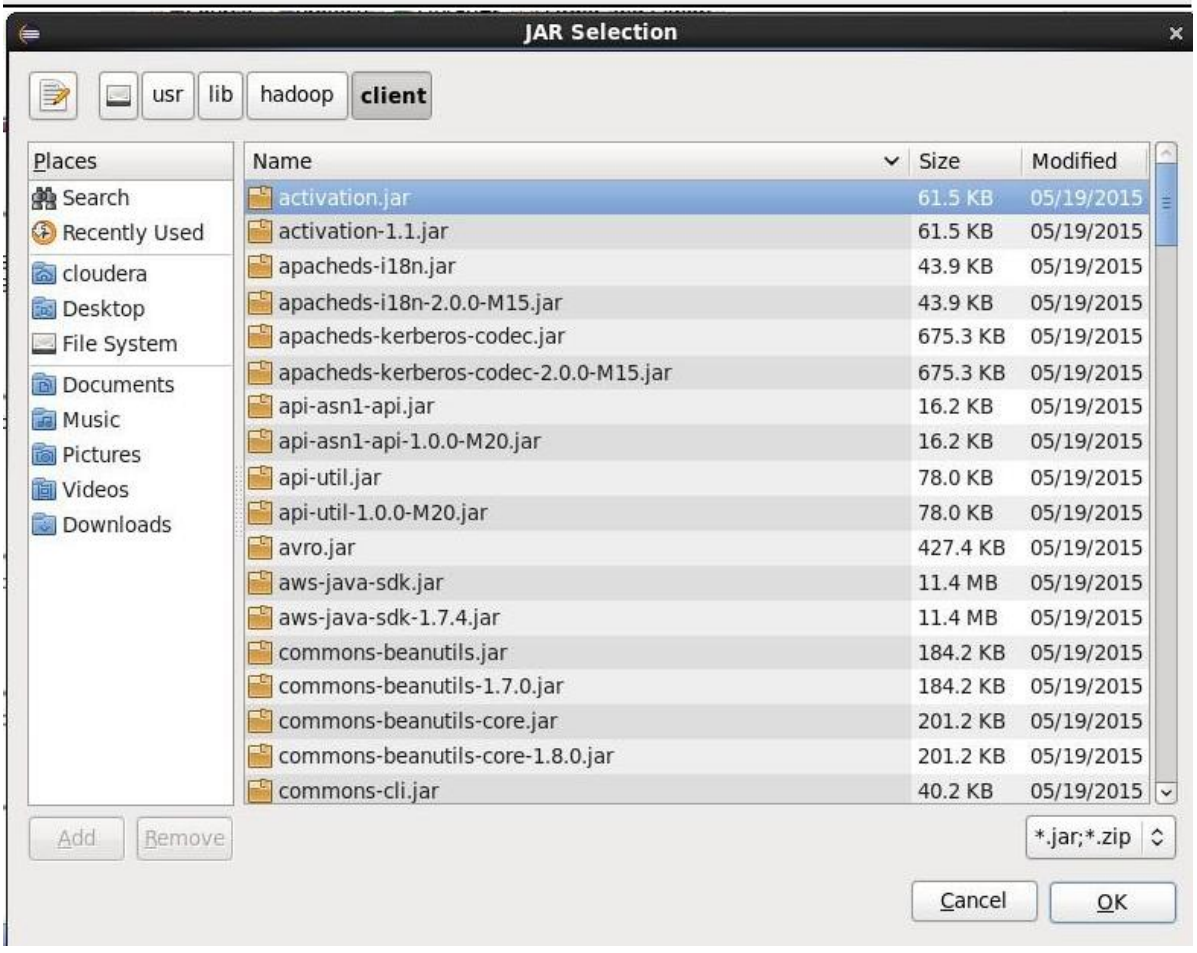

JAR Selection
usr
lib
Places
Search
Recently Used
cloudera
Desktop
File System
Documents
Music
Pictures
Videos
Downloads
Name
Size
Modified
anaconda-runtime 06/09/2015
avro 06/09/2015
bigtop-tomcat 06/09/2015
bigtop-utils 06/09/2015
bonobo 06/09/2015
ConsoleKit 06/09/2015
crunch 06/09/2015
cups 06/09/2015
dkms 06/09/2015
flume-ng 06/09/2015
games 09/23/2011
gcc 02/21/2013
hadoop 06/09/2015
hadoop-0.20-mapreduce 06/09/2015
hadoop-hdfs 06/09/2015
hadoop-httpfs 06/09/2015
hadoop-kms 06/09/2015
hadoop-mapreduce 06/09/2015
Add
Remove
.jar;.zip
Cancel
OK


JAR Selection
usr
lib
hadoop
Places
Search
Recently Used
cloudera
Desktop
File System
Documents
Music
Pictures
Videos
Downloads
Name
Size
Modified
bin 06/09/2015
client 06/09/2015
client-0.20 06/09/2015
cloudera 06/09/2015
etc 06/09/2015
lib 06/09/2015
libexec 06/09/2015
sbin 06/09/2015
hadoop-annotations.jar 21.0 KB 05/19/2015
hadoop-annotations-2.6.0-cdh5.4.2.jar 21.0 KB 05/19/2015
hadoop-auth.jar 71.4 KB 05/19/2015
hadoop-auth-2.6.0-cdh5.4.2.jar 71.4 KB 05/19/2015
hadoop-aws.jar 13.7 MB 05/19/2015
hadoop-aws-2.6.0-cdh5.4.2.jar 13.7 MB 05/19/2015
hadoop-common.jar 3.2 MB 05/19/2015
hadoop-common-2.6.0-cdh5.4.2.jar 3.2 MB 05/19/2015
hadoop-common-2.6.0-cdh5.4.2-tests.jar 1.7 MB 05/19/2015
hadoop-common-tests.jar 1.7 MB 05/19/2015
Add
Remove
.jar;.zip
Cancel
OK

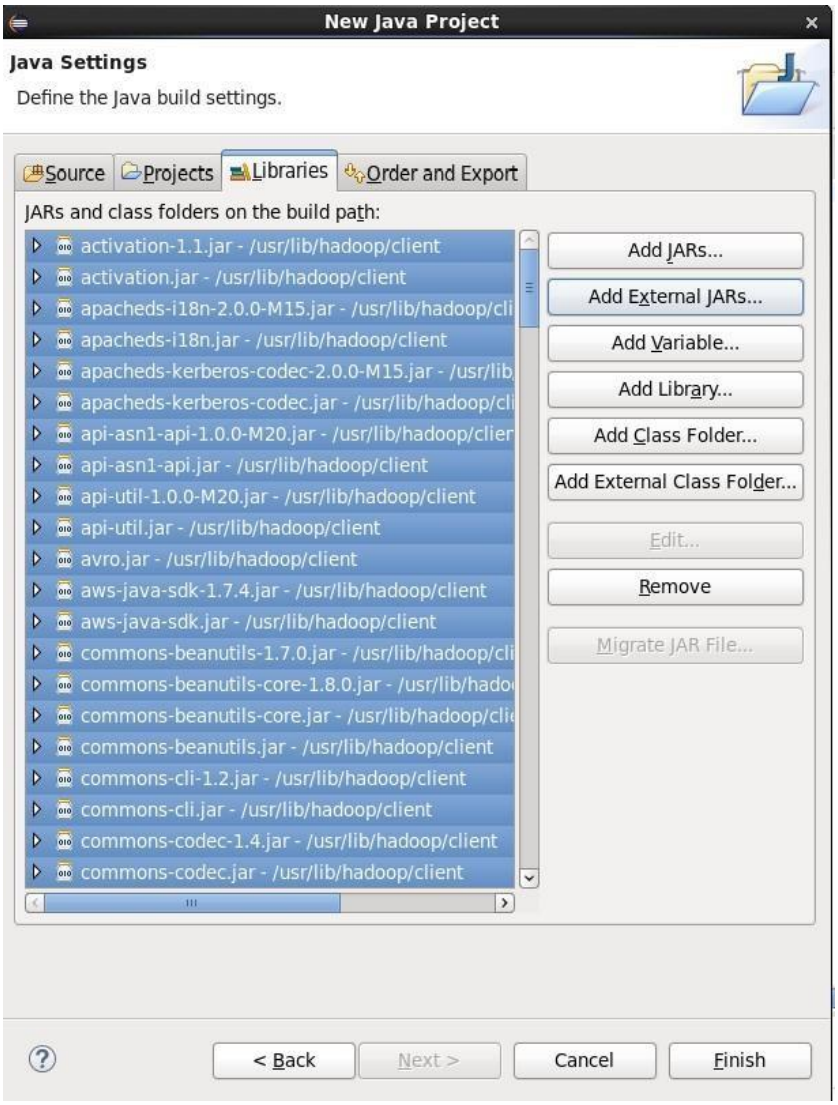
BIGDATA


New Java Project
Java Settings
Define the Java build settings.
Source
Projects
Libraries
Order and Export
JARs and class folders on the build path:
hadoop-annotations-2.6.0-cdh5.4.2.jar - /usr/lib/h
hadoop-annotations.jar - /usr/lib/hadoop
hadoop-auth-2.6.0-cdh5.4.2.jar - /usr/lib/hadoop
hadoop-auth.jar - /usr/lib/hadoop
hadoop-aws-2.6.0-cdh5.4.2.jar - /usr/lib/hadoop
hadoop-aws.jar - /usr/lib/hadoop
hadoop-common-2.6.0-cdh5.4.2-tests.jar - /usr/li
hadoop-common-2.6.0-cdh5.4.2.jar - /usr/lib/had
hadoop-common-tests.jar - /usr/lib/hadoop
hadoop-common.jar - /usr/lib/hadoop
hadoop-nfs-2.6.0-cdh5.4.2.jar - /usr/lib/hadoop
hadoop-nfs.jar - /usr/lib/hadoop
parquet-avro.jar - /usr/lib/hadoop
parquet-cascading.jar - /usr/lib/hadoop
parquet-column.jar - /usr/lib/hadoop
parquet-common.jar - /usr/lib/hadoop
parquet-encoding.jar - /usr/lib/hadoop
parquet-format-javadoc.jar - /usr/lib/hadoop
parquet-format-sources.jar - /usr/lib/hadoop
parquet-format.jar - /usr/lib/hadoop
parquet-generator.jar - /usr/lib/hadoop
Add JARs...
Add External JARs...
Add Variable...
Add Library...
Add Class Folder...
Add External Class Folder...
Edit...
Remove
Migrate JAR File...
< Back
Next >
Cancel
Finish


JAR Selection
usr
lib
hadoop
Places
Search
Recently Used
cloudera
Desktop
File System
Documents
Music
Pictures
Videos
Downloads
Name
Size
Modified
bin 06/09/2015
client 06/09/2015
client-0.20 06/09/2015
cloudera 06/09/2015
etc 06/09/2015
lib 06/09/2015
libexec 06/09/2015
sbin 06/09/2015
hadoop-annotations.jar 21.0 KB 05/19/2015
hadoop-annotations-2.6.0-cdh5.4.2.jar 21.0 KB 05/19/2015
hadoop-auth.jar 71.4 KB 05/19/2015
hadoop-auth-2.6.0-cdh5.4.2.jar 71.4 KB 05/19/2015
hadoop-aws.jar 13.7 MB 05/19/2015
hadoop-aws-2.6.0-cdh5.4.2.jar 13.7 MB 05/19/2015
hadoop-common.jar 3.2 MB 05/19/2015
hadoop-common-2.6.0-cdh5.4.2.jar 3.2 MB 05/19/2015
hadoop-common-2.6.0-cdh5.4.2-tests.jar 1.7 MB 05/19/2015
hadoop-common-tests.jar 1.7 MB 05/19/2015
Add
Remove
.jar;.zip
Cancel
OK


JAR Selection
usr
lib
hadoop
client
Places
Search
Recently Used
cloudera
Desktop
File System
Documents
Music
Pictures
Videos
Downloads
Name
Size
Modified
activation.jar 61.5 KB 05/19/2015
activation-1.1.jar 61.5 KB 05/19/2015
apacheds-i18n.jar 43.9 KB 05/19/2015
apacheds-i18n-2.0.0-M15.jar 43.9 KB 05/19/2015
apacheds-kerberos-codec.jar 675.3 KB 05/19/2015
apacheds-kerberos-codec-2.0.0-M15.jar 675.3 KB 05/19/2015
api-asn1-api.jar 16.2 KB 05/19/2015
api-asn1-api-1.0.0-M20.jar 16.2 KB 05/19/2015
api-util.jar 78.0 KB 05/19/2015
api-util-1.0.0-M20.jar 78.0 KB 05/19/2015
avro.jar 427.4 KB 05/19/2015
aws-java-sdk.jar 11.4 MB 05/19/2015
aws-java-sdk-1.7.4.jar 11.4 MB 05/19/2015
commons-beanutils.jar 184.2 KB 05/19/2015
commons-beanutils-1.7.0.jar 184.2 KB 05/19/2015
commons-beanutils-core.jar 201.2 KB 05/19/2015
commons-beanutils-core-1.8.0.jar 201.2 KB 05/19/2015
commons-cli.jar 40.2 KB 05/19/2015
Add
Remove
.jar;.zip
Cancel
OK

BIGDATA



StubDriver.java | Union.java | Intersection.java | MatrixMultiplicationMapper.java | MatrixMultiplicationReducer.java | MatrixMultiplicatic

```java
package demo;

import org.apache.hadoop.io.Text;
import java.io.IOException;
import java.util.HashMap;

public class MatrixMultiplicationReducer extends org.apache.hadoop.mapreduce.Reducer<Text, Text, Text, Text> {

    @Override
    public void reduce(Text key, Iterable<Text> values, Context context)
            throws IOException, InterruptedException {

        // key = (i,k)
        // values = [(M/N,j,V/W), ...]
        HashMap<Integer, Float> hashB = new HashMap<Integer, Float>();

        // Separate M and N values
        for (Text val : values) {
            String[] value = val.toString().split(",");
            if (value[0].equals("M")) {
                hashB.put(Integer.parseInt(value[1]), Float.parseFloat(value[2]));
            } else if (value[0].equals("N")) {
                hashB.put(Integer.parseInt(value[1]), Float.parseFloat(value[2]));
            }
        }

        int n = Integer.parseInt(context.getConfiguration().get("n"));
        float result = 0.0f;
        // Multiply and sum over matching indices j
        for (int j = 0; j < n; j++) {
            float m_ij = hashB.containsKey(j) ? hashB.get(j) : 0.0f;
            float n_jk = hashB.containsKey(j) ? hashB.get(j) : 0.0f;
            result += m_ij * n_jk;
        }

        // Emit only non-zero results
        if (result != 0.0f) {
            context.write(null, new Text(key.toString() + "," + Float.toString(result)));
        }
    }
}
```

BIGDATA

StubDriver.java | Union.java | Intersection.java | MatrixMultiplicationMapper.java | MatrixMultip

```java
package demo;

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

public class MatrixMultiplicationDriver {

    public static void main(String[] args) throws Exception {

        if (args.length != 2) {
            System.err.println("Usage: MatrixMultiplication <input_dir> <output_dir>");
            System.exit(2);
        }

        Configuration conf = new Configuration();
        // M is an m×n matrix; N is an n×p matrix
        conf.set("m", "1000");  // rows in M
        conf.set("n", "100");   // columns in M (and rows in N)
        conf.set("p", "1000");  // columns in N

        Job job = Job.getInstance(conf, "Matrix Multiplication");

        job.setJarByClass(MatrixMultiplicationDriver.class);
        job.setMapperClass(MatrixMultiplicationMapper.class);
        job.setReducerClass(MatrixMultiplicationReducer.class);
        job.setOutputKeyClass(Text.class);
        job.setOutputValueClass(Text.class);

        job.setInputFormatClass(TextInputFormat.class);
        job.setOutputFormatClass(TextOutputFormat.class);

        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputPath(job, new Path(args[1]));

        System.exit(job.waitForCompletion(true) ? 0 : 1);
    }
}
```

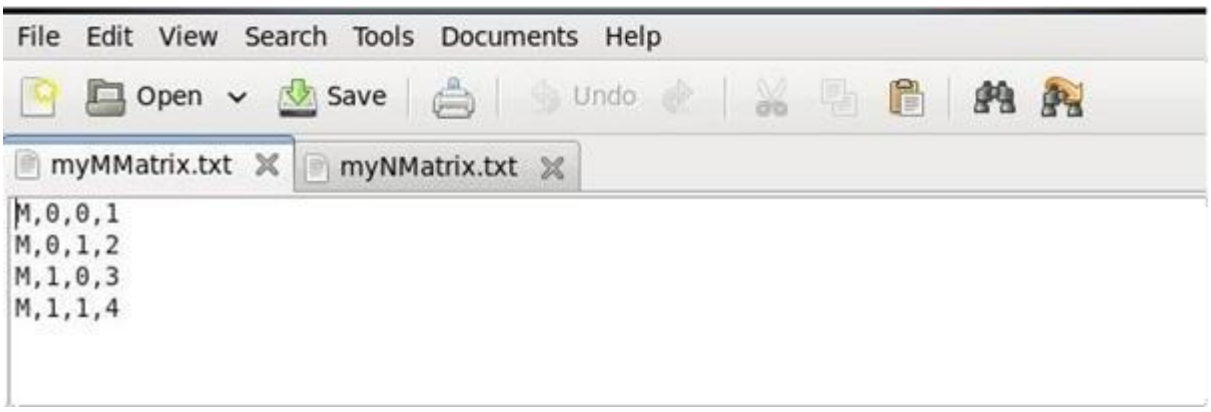
StubDriver.java | Union.java | Intersection.java | MatrixMultiplicationMapper.java | MatrixMultiplicationReduc

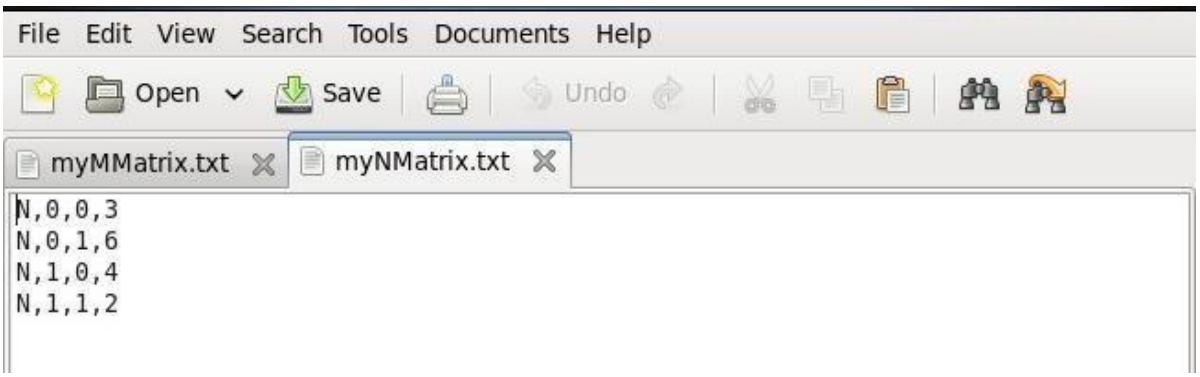
```java
package demo;

import java.io.IOException;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
public class MatrixMultiplicationMapper extends Mapper<LongWritable, Text, Text, Text> {

    @Override
    public void map(LongWritable key, Text value, Context context) throws IOException, InterruptedException {

        Configuration conf = context.getConfiguration();

        int m = Integer.parseInt(conf.get("m"));
        int p = Integer.parseInt(conf.get("p"));

        String line = value.toString();
        String[] indicesAndValue = line.split(",");
        Text outputKey = new Text();
        Text outputValue = new Text();

        // Input format examples:
        // M,i,j,Mij or N,j,k,Njk

        if (indicesAndValue[0].equals("M")) {
            // For matrix M(i,j), emit (i,k) as key and (M,j,Mij) as value for all k
            for (int k = 0; k < p; k++) {
                outputKey.set(indicesAndValue[1] + "," + k);
                outputValue.set("M," + indicesAndValue[2] + "," + indicesAndValue[3]);  // value = (M,j,Mij)
                context.write(outputKey, outputValue);
            }
        } else if (indicesAndValue[0].equals("N")) {
            // For matrix N(j,k), emit (i,k) as key and (N,j,Njk) as value for all i
            for (int i = 0; i < m; i++) {
                outputKey.set(i + "," + indicesAndValue[2]);
                outputValue.set("N," + indicesAndValue[1] + "," + indicesAndValue[3]);  // value = (N,j,Njk)
                context.write(outputKey, outputValue);
            }
        }
    }
}
```

BIGDATA

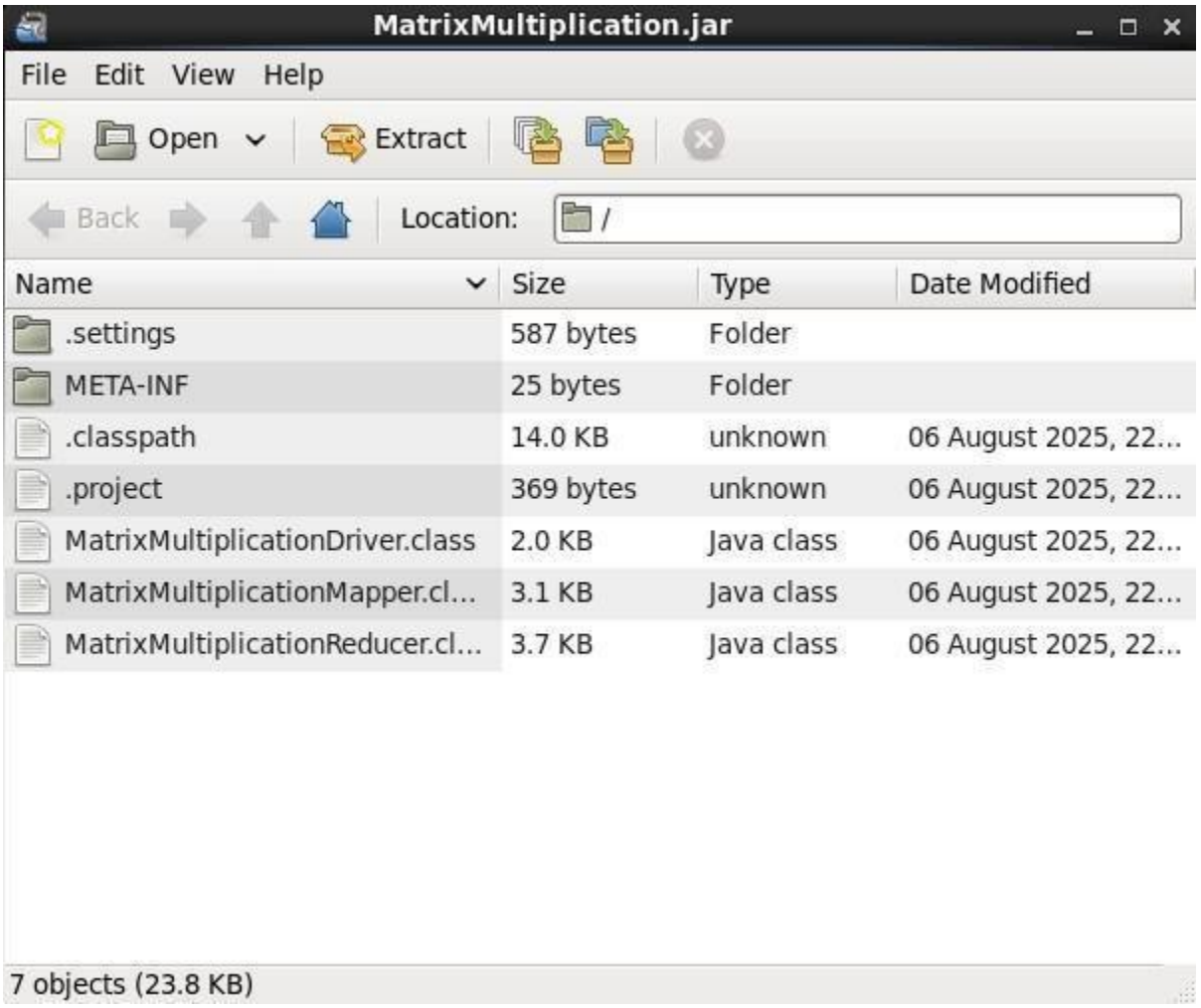

cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ ls
express-deployment.json
localfile.txt-
MatrixMultiplication.jar
myInputFile.4.txt
myInputFile.txt~
WordCount.jar
[cloudera@quickstart ~]$ hdfs dfs -put MatrixMultiplication.jar /user/cloudera
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 7 items
drwx------ - cloudera cloudera 0 2025-08-06 22:00 .Trash
drwx------ - cloudera cloudera 0 2025-08-06 21:44 .staging
-rw-r--r-- 1 cloudera cloudera 6555 2025-08-06 23:08 MatrixMultiplication.jar
-rw-r--r-- 1 cloudera cloudera 4887 2025-08-05 23:20 WordCount.jar
drwxr-xr-x - cloudera cloudera 0 2025-07-30 00:31 manas
-rw-r--r-- 1 cloudera cloudera 52 2025-08-05 23:25 myInputFile.txt
drwxr-xr-x - cloudera cloudera 0 2025-08-06 21:43 myOutput
[cloudera@quickstart ~]$ cat > myMMatrix.txt
M,0,0,1
M,0,1,2
M,1,0,3
M,1,1,4
^Z
[1]+ Stopped cat > myMMatrix.txt
[cloudera@quickstart ~]$ cat myMMatrix.txt
M,0,0,1
M,0,1,2
M,1,0,3
M,1,1,4
[cloudera@quickstart ~]$ cat > myNMatrix.txt
N,0,0,3
N,0,1,6
N,1,0,4
N,1,1,2
^Z
[2]+ Stopped cat > myNMatrix.txt
[cloudera@quickstart ~]$ cat myNMatrix.txt
N,0,0,3
N,0,1,6
N,1,0,4
N,1,1,2


[cloudera@quickstart ~]$ hdfs dfs -mkdir /user/cloudera/matrixInput
[cloudera@quickstart ~]$ hdfs dfs -put myMMatrix.txt /user/cloudera/matrixInput/
[cloudera@quickstart ~]$ hdfs dfs -put myNMatrix.txt /user/cloudera/matrixInput/
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 8 items
drwx------ - cloudera cloudera 0 2025-08-06 22:00 .Trash
drwx------ - cloudera cloudera 0 2025-08-06 21:44 .staging
-rw-r--r-- 1 cloudera cloudera 6555 2025-08-06 23:08 MatrixMultiplication.jar
-rw-r--r-- 1 cloudera cloudera 4887 2025-08-05 23:20 WordCount.jar
drwxr-xr-x - cloudera cloudera 0 2025-07-30 00:31 manas
drwxr-xr-x - cloudera cloudera 0 2025-08-06 23:14 matrixInput
-rw-r--r-- 1 cloudera cloudera 52 2025-08-05 23:25 myInputFile.txt
drwxr-xr-x - cloudera cloudera 0 2025-08-06 21:43 myOutput
[cloudera@quickstart ~]$ hadoop jar MatrixMultiplication.jar MatrixMultiplicationDriver matrixInput matrixOutput
25/08/06 23:15:43 INFO client.RMProxy: Connecting to ResourceManager at quickstart.cloudera/10.0.2.15:8032
25/08/06 23:15:44 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
25/08/06 23:15:45 INFO input.FileInputFormat: Total input paths to process : 2
25/08/06 23:15:45 INFO mapreduce.JobSubmitter: number of splits:2
25/08/06 23:15:45 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1754541476929_0002
25/08/06 23:15:46 INFO impl.YarnClientImpl: Submitted application application_1754541476929_0002
25/08/06 23:15:46 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1754541476929_0002/
25/08/06 23:15:46 INFO mapreduce.Job: Running job: job_1754541476929_0002
25/08/06 23:16:03 INFO mapreduce.Job: Job job_1754541476929_0002 running in uber mode : false
25/08/06 23:16:03 INFO mapreduce.Job: map 0% reduce 0%
25/08/06 23:16:23 INFO mapreduce.Job: map 100% reduce 0%
25/08/06 23:16:34 INFO mapreduce.Job: map 100% reduce 100%
25/08/06 23:16:35 INFO mapreduce.Job: Job job_1754541476929_0002 completed successfully
25/08/06 23:16:35 INFO mapreduce.Job: Counters: 49


File Edit View Search Tools Documents Help
Open
Save
Undo
myMMatrix.txt
myNMatrix.txt
M,0,0,1
M,0,1,2
M,1,0,3
M,1,1,4

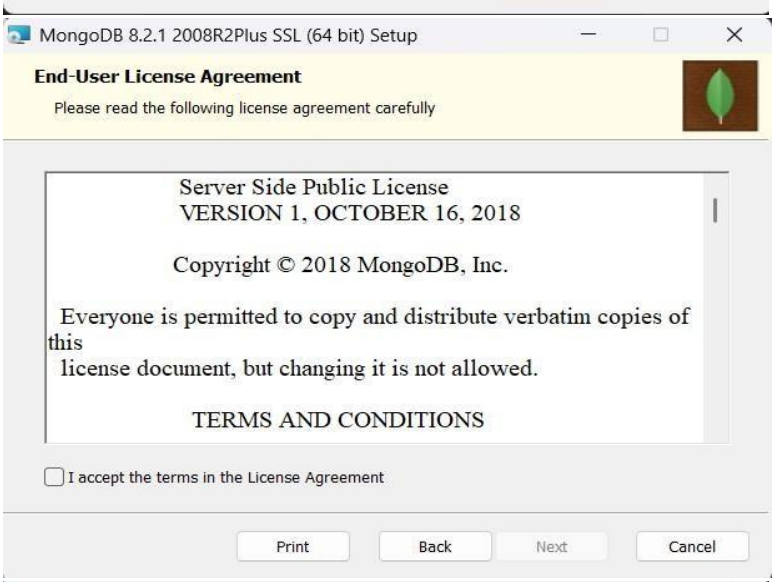

File Edit View Search Tools Documents Help
Open
Save
Undo
myMMatrix.txt
myNMatrix.txt
N,0,0,3
N,0,1,6
N,1,0,4
N,1,1,2

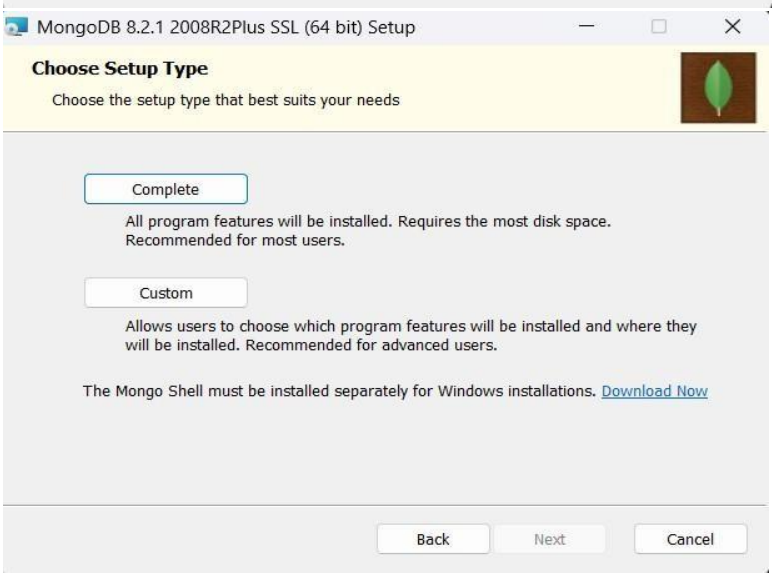
BIGDATA


MatrixMultiplication.jar
File Edit View Help
Open
Extract
Back
Location: /
Name
Size
Type
Date Modified
.settings 587 bytes Folder
META-INF 25 bytes Folder
.classpath 14.0 KB unknown 06 August 2025, 22...
.project 369 bytes unknown 06 August 2025, 22...
MatrixMultiplicationDriver.class 2.0 KB Java class 06 August 2025, 22...
MatrixMultiplicationMapper.cl... 3.1 KB Java class 06 August 2025, 22...
MatrixMultiplicationReducer.cl... 3.7 KB Java class 06 August 2025, 22...
7 objects (23.8 KB)

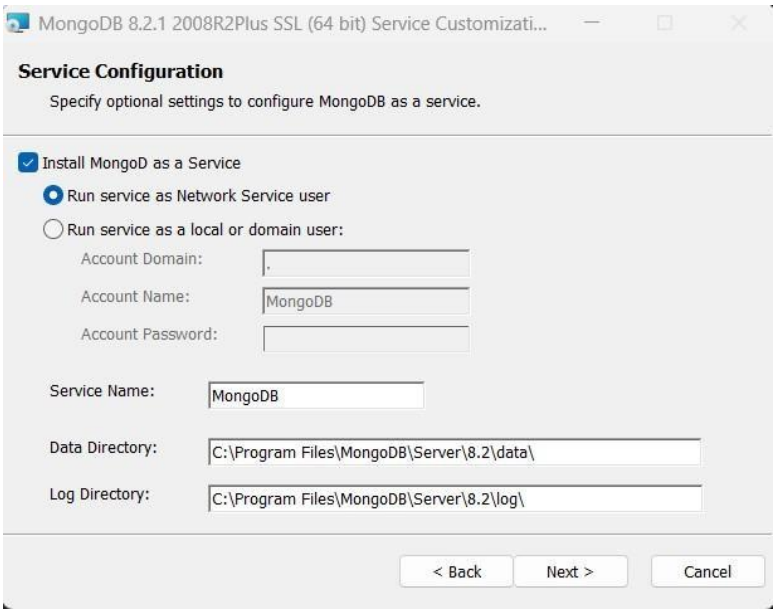
3. MongoDB

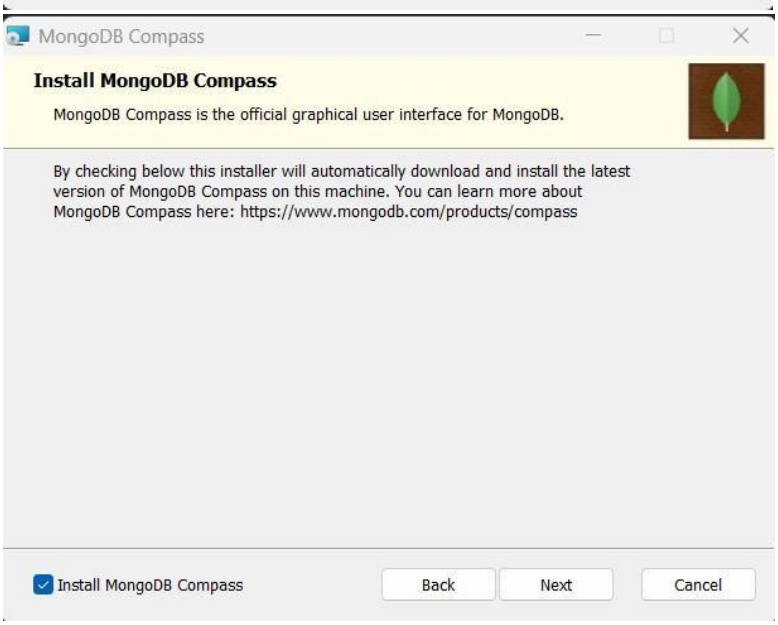
Installation

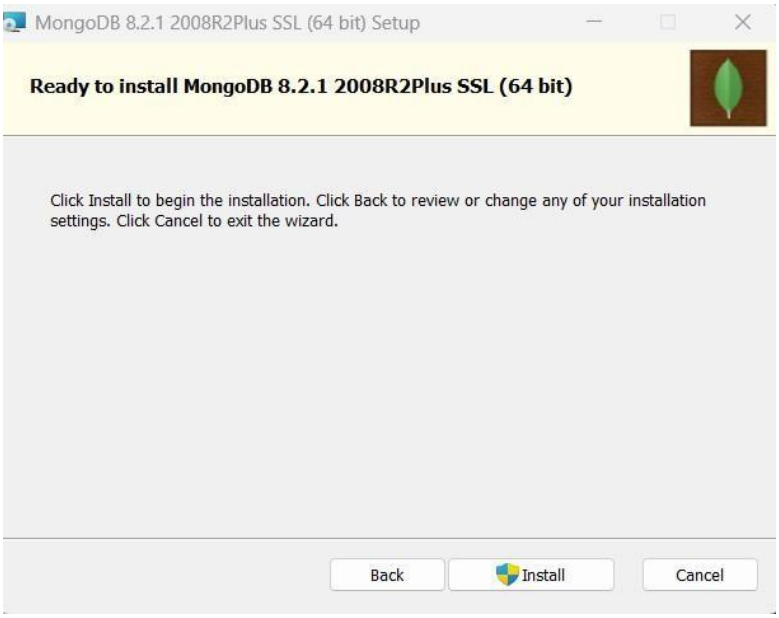




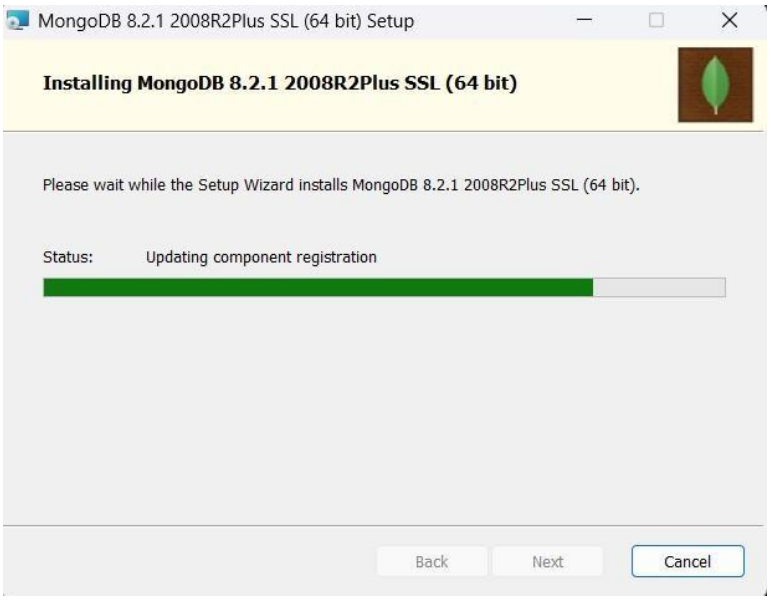
BIGDATA

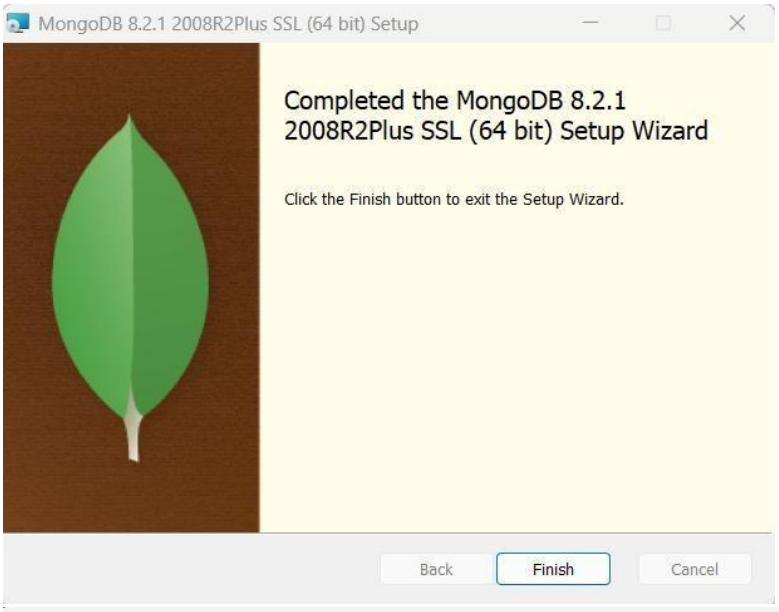

MongoDB 8.2.1 2008R2Plus SSL (64 bit) Service Customizati...
Service Configuration
Specify optional settings to configure MongoDB as a service.
Install MongoD as a Service
Run service as Network Service user
Run service as a local or domain user:
Account Domain:
.
Account Name:
MongoDB
Account Password:
Service Name:
MongoDB
Data Directory:
C:\Program Files\MongoDB\Server\8.2\data\
Log Directory:
C:\Program Files\MongoDB\Server\8.2\log\
< Back
Next >
Cancel

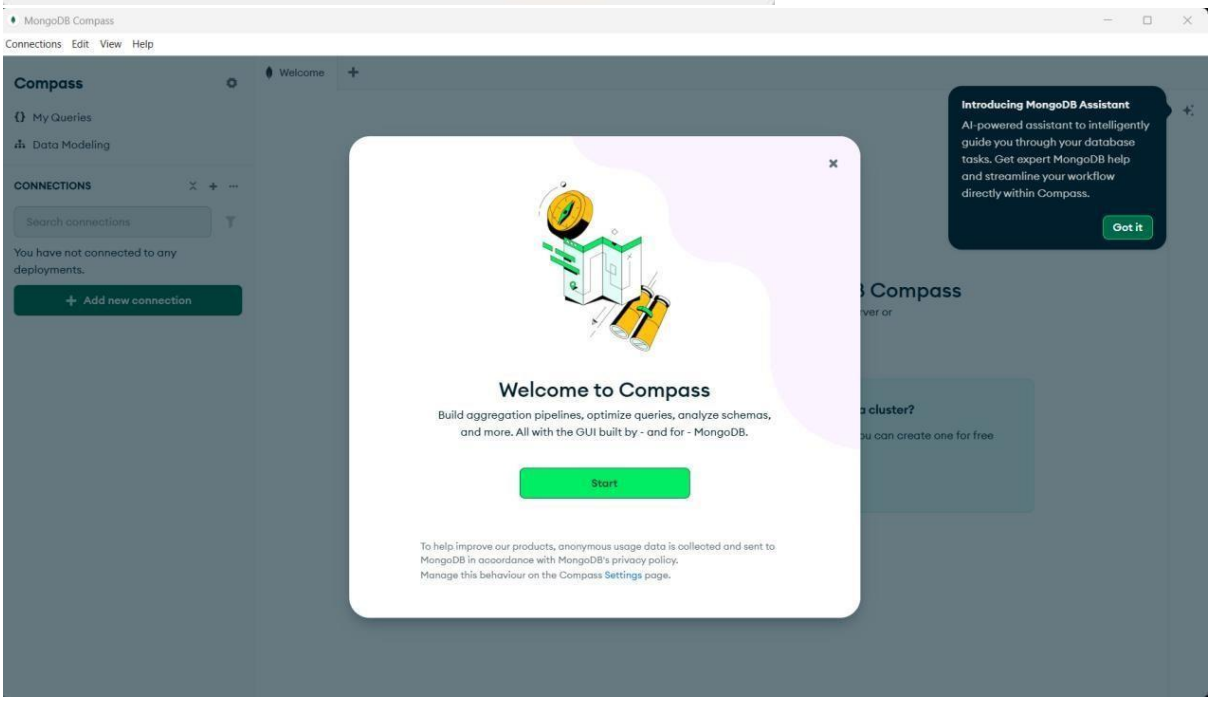

MongoDB Compass
Install MongoDB Compass
MongoDB Compass is the official graphical user interface for MongoDB.
By checking below this installer will automatically download and install the latest version of MongoDB Compass on this machine. You can learn more about MongoDB Compass here: https://www.mongodb.com/products/compass
Install MongoDB Compass
Back
Next
Cancel


MongoDB 8.2.1 2008R2Plus SSL (64 bit) Setup
Ready to install MongoDB 8.2.1 2008R2Plus SSL (64 bit)
Click Install to begin the installation. Click Back to review or change any of your installation settings. Click Cancel to exit the wizard.
Back
Install
Cancel

BIGDATA


MongoDB 8.2.1 2008R2Plus SSL (64 bit) Setup
Installing MongoDB 8.2.1 2008R2Plus SSL (64 bit)
Please wait while the Setup Wizard installs MongoDB 8.2.1 2008R2Plus SSL (64 bit).
Status: Updating component registration
Back
Next
Cancel


MongoDB 8.2.1 2008R2Plus SSL (64 bit) Setup
Completed the MongoDB 8.2.1 2008R2Plus SSL (64 bit) Setup Wizard
Click the Finish button to exit the Setup Wizard.
Back
Finish
Cancel


MongoDB Compass
Connections Edit View Help
Compass
My Queries
Data Modeling
CONNECTIONS
Search connections
You have not connected to any deployments.
Add new connection
Welcome
Introducing MongoDB Assistant
AI-powered assistant to intelligently guide you through your database tasks. Get expert MongoDB help and streamline your workflow directly within Compass.
Got it
Welcome to Compass
Build aggregation pipelines, optimize queries, analyze schemas, and more. All with the GUI built by - and for - MongoDB.
Start
To help improve our products, anonymous usage data is collected and sent to MongoDB in accordance with MongoDB's privacy policy.
Manage this behaviour on the Compass Settings page.

Sample Database Creation

“The use Command” is used to create database. The command will create a new database if it doesn't exist, otherwise it will return the existing database.

```
test> use sampleDB
switched to db sampleDB
sampleDB>
```

```
sampleDB> db.employee.insert({
... "empcode":123,
... "empfname":"iron",
... "emplname":"man",
... "job":"Engineer",
... "hiredate":"1985-12-15",
... "salary":2000
... })
DeprecationWarning: Collection.insert() is deprecated. Use insertOne, insertMany, or bulkWrite.
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('6915b616a4e9dff14663b112') }
}
sampleDB> db.employee.insertOne({ "empcode": 123, "empfname": "iron", "emplname": "man", "job": "Engineer", "hiredate":
"1985-12-15", "salary": 2000 })
{
  acknowledged: true,
  insertedId: ObjectId('6915b66ea4e9dff14663b113')
}
```

```
sampleDB> db.employee.insertMany([{ "empcode": 124, "empfname": "power", "emplname": "house", "job": "hr", "hiredate": "
1992-5-18", "salary": 2000 }, { "empcode": 125, "empfname": "tony", "emplname": "pawar", "job": "hr", "hiredate": "1992-
5-18", "salary": 2000 }, { "empcode": 126, "empfname": "captain", "emplname": "raut", "job": "Engineer", "hiredate": "19
92-5-18", "salary": 2078 }, { "empcode": 127, "empfname": "captain", "emplname": "america", "job": "Engineer", "hiredate
": "1992-5-18", "salary": 2078 }]);
{
  acknowledged: true,
  insertedIds: {
    '0': ObjectId('6915b95ea4e9dff14663b114'),
    '1': ObjectId('6915b95ea4e9dff14663b115'),
    '2': ObjectId('6915b95ea4e9dff14663b116'),
    '3': ObjectId('6915b95ea4e9dff14663b117')
  }
}
```


Query the Sample Database using MongoDB querying command

BIGDATA

```
sampleDB> db.employee.find()
[
  {
    _id: ObjectId('6915b616a4e9dff14663b112'),
    empcode: 123,
    empfname: 'iron',
    emplname: 'man',
    job: 'Engineer',
    hiredate: '1985-12-15',
    salary: 2000
  },
  {
    _id: ObjectId('6915b66ea4e9dff14663b113'),
    empcode: 123,
    empfname: 'iron',
    emplname: 'man',
    job: 'Engineer',
    hiredate: '1985-12-15',
    salary: 2000
  },
  {
    _id: ObjectId('6915b95ea4e9dff14663b114'),
    empcode: 124,
    empfname: 'power',
    emplname: 'house',
    job: 'hr',
    hiredate: '1992-5-18',
    salary: 2000
  },
  {
    _id: ObjectId('6915b95ea4e9dff14663b115'),
    empcode: 125,
    empfname: 'tony',
    emplname: 'pawar',
    job: 'hr',
    hiredate: '1992-5-18',
    salary: 2000
  },
```

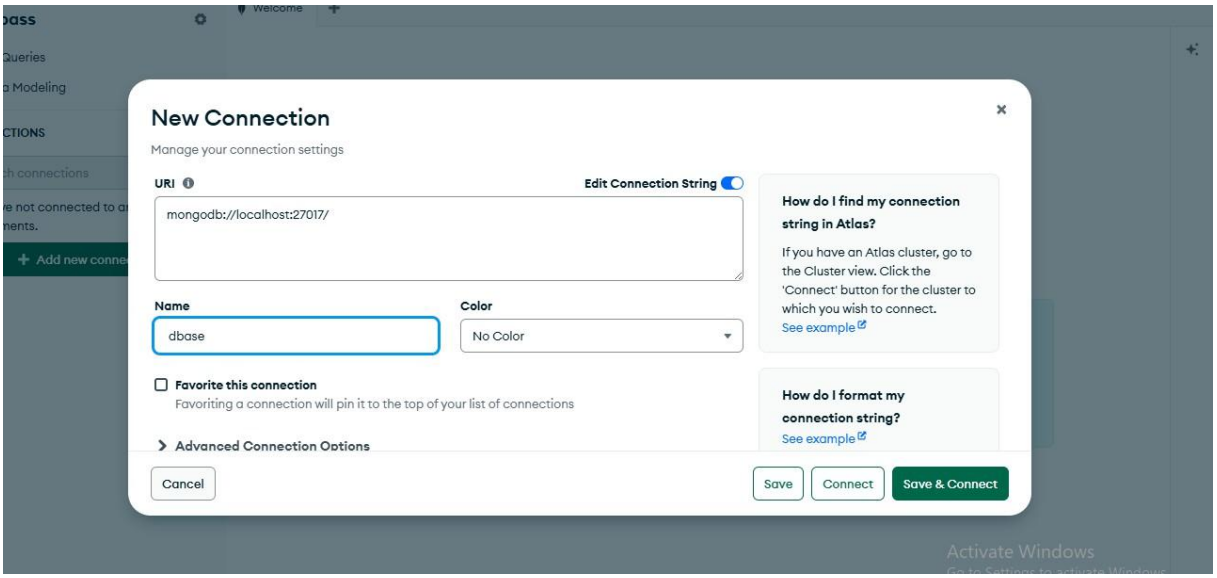
```
sampleDB> db.employee.findOne({empfname:"power"})
{
  _id: ObjectId('6915b95ea4e9dff14663b114'),
  empcode: 124,
  empfname: 'power',
  emplname: 'house',
  job: 'hr',
  hiredate: '1992-5-18',
  salary: 2000
}
```

BIGDATA

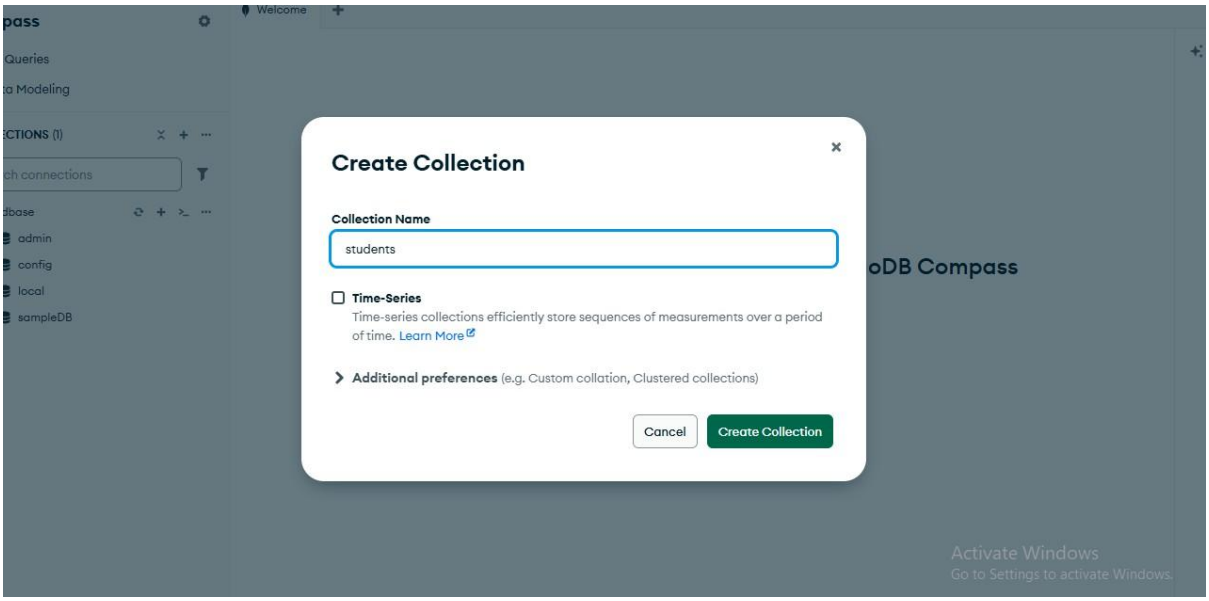
```
sampleDB> db.employee.find({job:"Engineer"})
[
  {
    _id: ObjectId('6915b616a4e9dff14663b112'),
    empcode: 123,
    empfname: 'iron',
    emplname: 'man',
    job: 'Engineer',
    hiredate: '1985-12-15',
    salary: 2000
  },
  {
    _id: ObjectId('6915b66ea4e9dff14663b113'),
    empcode: 123,
    empfname: 'iron',
    emplname: 'man',
    job: 'Engineer',
    hiredate: '1985-12-15',
    salary: 2000
  },
  {
    _id: ObjectId('6915b95ea4e9dff14663b116'),
    empcode: 126,
    empfname: 'captain',
    emplname: 'raut',
    job: 'Engineer',
    hiredate: '1992-5-18',
    salary: 2078
  },
  {
    _id: ObjectId('6915b95ea4e9dff14663b117'),
    empcode: 127,
    empfname: 'captain',
    emplname: 'america',
    job: 'Engineer',
    hiredate: '1992-5-18',
    salary: 2078
  }
]
```

```
sampleDB> db.employee.find({salary: {$gt: 2000}})
[
  {
    _id: ObjectId('6915b95ea4e9dff14663b116'),
    empcode: 126,
    empfname: 'captain',
    emplname: 'raut',
    job: 'Engineer',
    hiredate: '1992-5-18',
    salary: 2078
  },
  {
    _id: ObjectId('6915b95ea4e9dff14663b117'),
    empcode: 127,
    empfname: 'captain',
    emplname: 'america',
    job: 'Engineer',
    hiredate: '1992-5-18',
    salary: 2078
  }
]
```

Operations on the collection with MongoDB Compass GUI
For that we have to open MongoDB Compass & create the new connection with following values and make sure to click on Save and connect

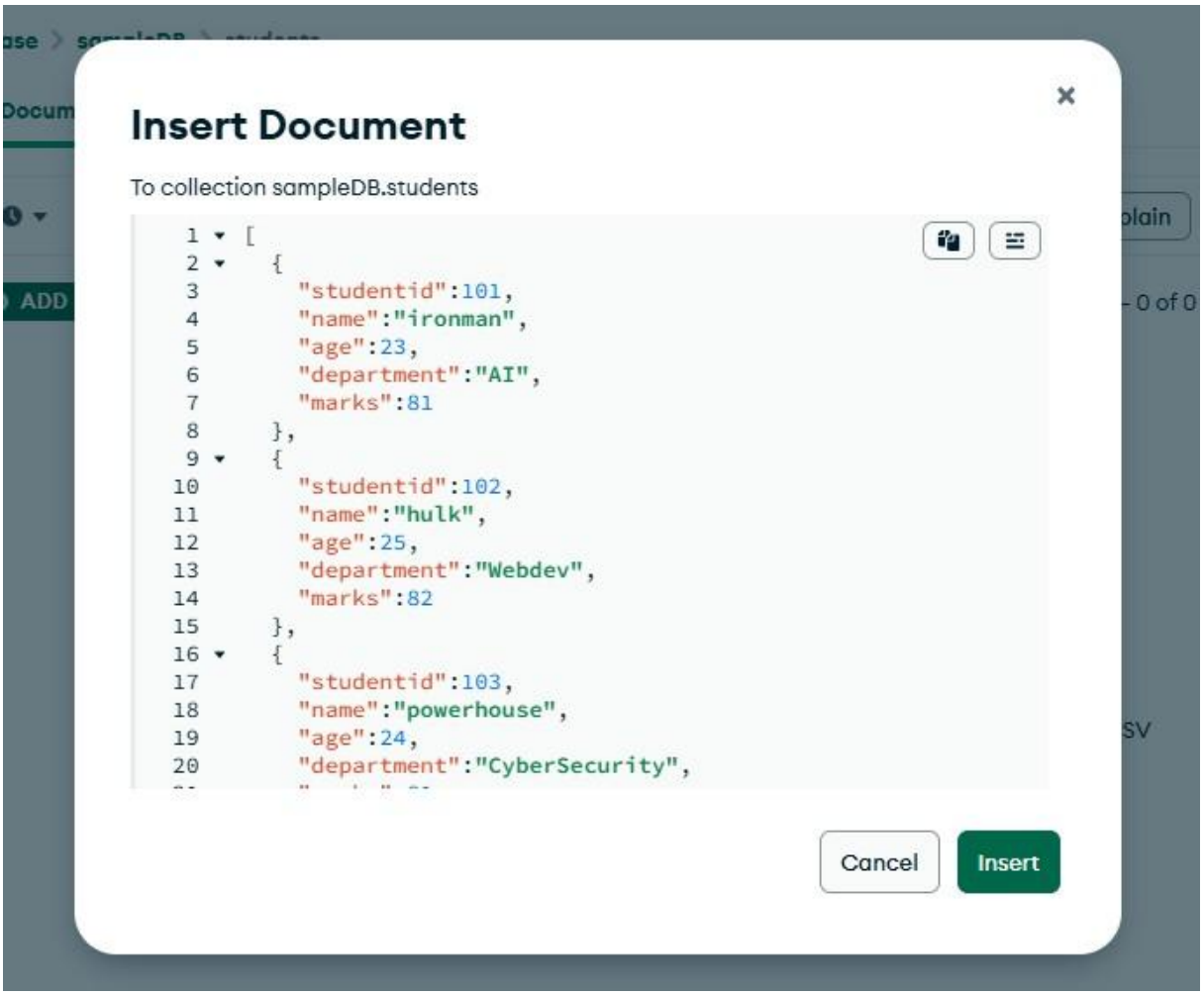


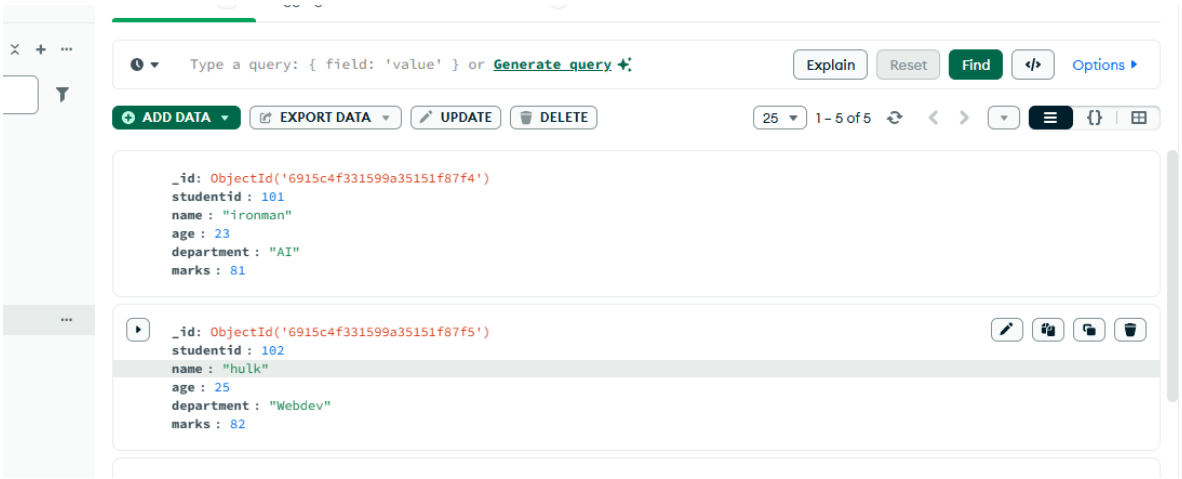
Create Collection



BIGDATA

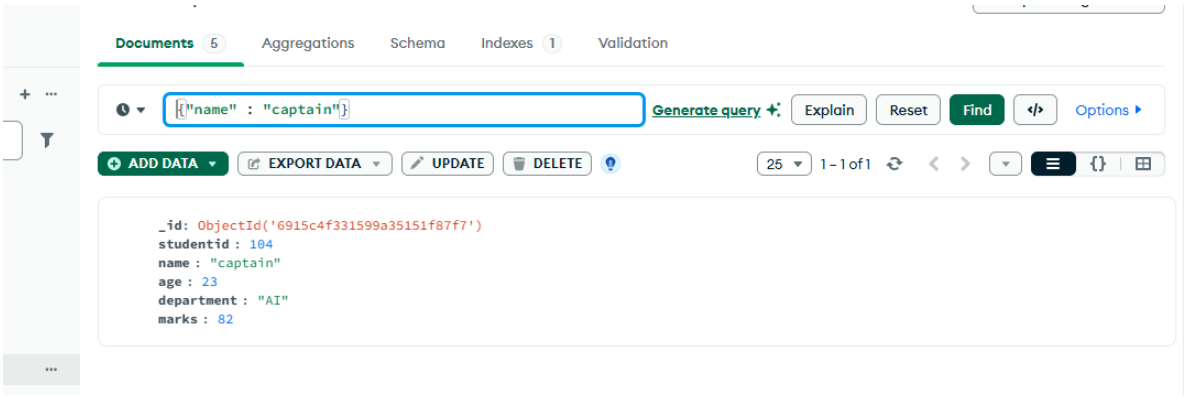
Insert Document

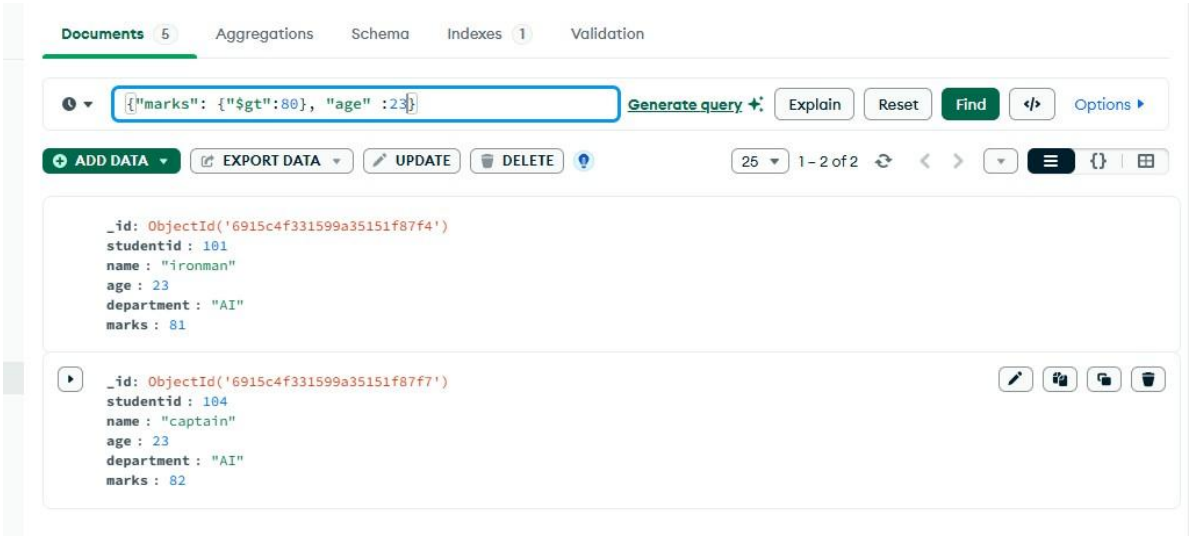




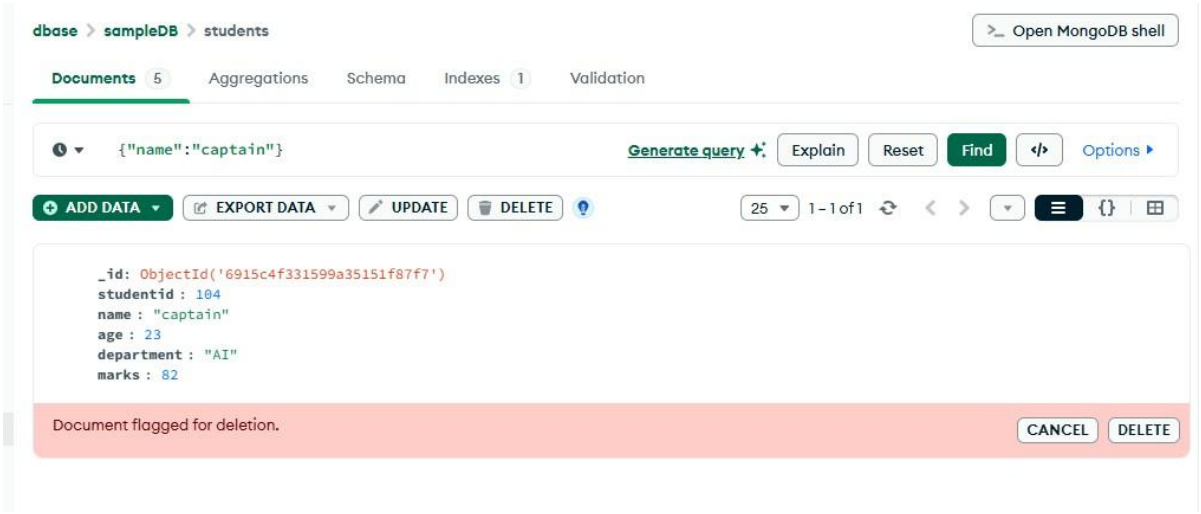
BIGDATA

Query Document





Delete Document



Indexing

```
> use sampleDB
< switched to db sampleDB
> db.students.createIndex({studentid: 101})
< studentid_101
> db.students.getIndexes()
< [
    { v: 2, key: { _id: 1 }, name: '_id_' },
    { v: 2, key: { studentid: 101 }, name: 'studentid_101' }
  ]
sampleDB>
```

4. Hive

1. Hive Data Types

Primitive Data Types:

• Numeric Types:

• TINYINT: 1-byte signed integer.

• SMALLINT: 2-byte signed integer.

• INT / INTEGER: 4-byte signed integer.

• BIGINT: 8-byte signed integer.

• FLOAT: 4-byte single-precision floating-point number.

• DOUBLE: 8-byte double-precision floating-point number.

• DECIMAL: Fixed-point decimal numbers with user-defined precision and scale.

• Date/Time Types:

• DATE: Represents a date in 'YYYY-MM-DD' format.

• TIMESTAMP: Represents a point in time with date and time, including optional

nanosecond precision.

• INTERVAL: Represents a period of time.

• String Types:

• STRING: Variable-length character string.

• VARCHAR(n): Variable-length character string with a maximum length of 'n'.

• CHAR(n): Fixed-length character string with a length of 'n', padded with spaces if shorter.

• Miscellaneous Types:

• BOOLEAN: Stores TRUE or FALSE values.

• BINARY: Stores a sequence of bytes.

• Complex Data Types:

• ARRAY<data_type>:

An ordered collection of elements of the same data_type. Similar to a list in other programming languages.

• MAP<key_data_type, value_data_type>:

An associative array (key-value pairs) where keys are unique and of key_data_type, and values are of value_data_type

• STRUCT<col_name : data_type, ...>:

A record type that groups together fields (columns) of potentially different data types, similar to a struct in C or a record in other systems.

• UNIONTYPE<data_type1, data_type2, ...>:

A type that can hold a value of one of several specified data types at any given time.

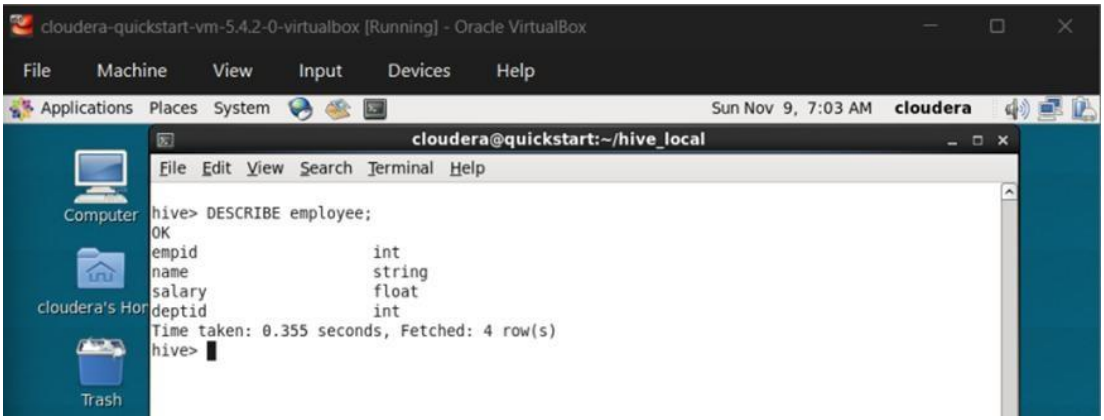
Create Database & Table in Hive
show databases - To check default database provided by Hive

create database:- To create a new database

Create table employee

BIGDATA

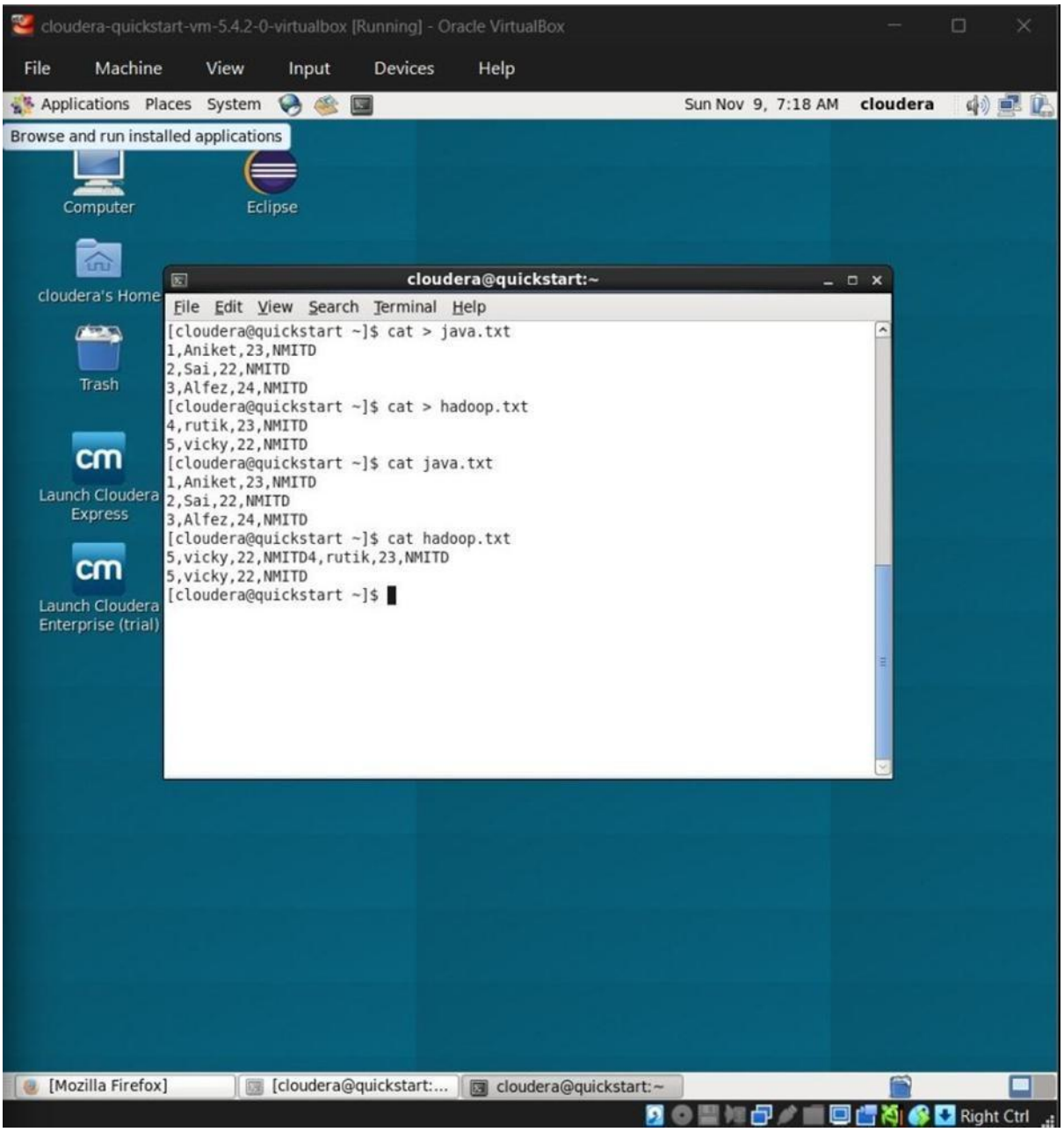
Describe the table employee



Hive Partitioning

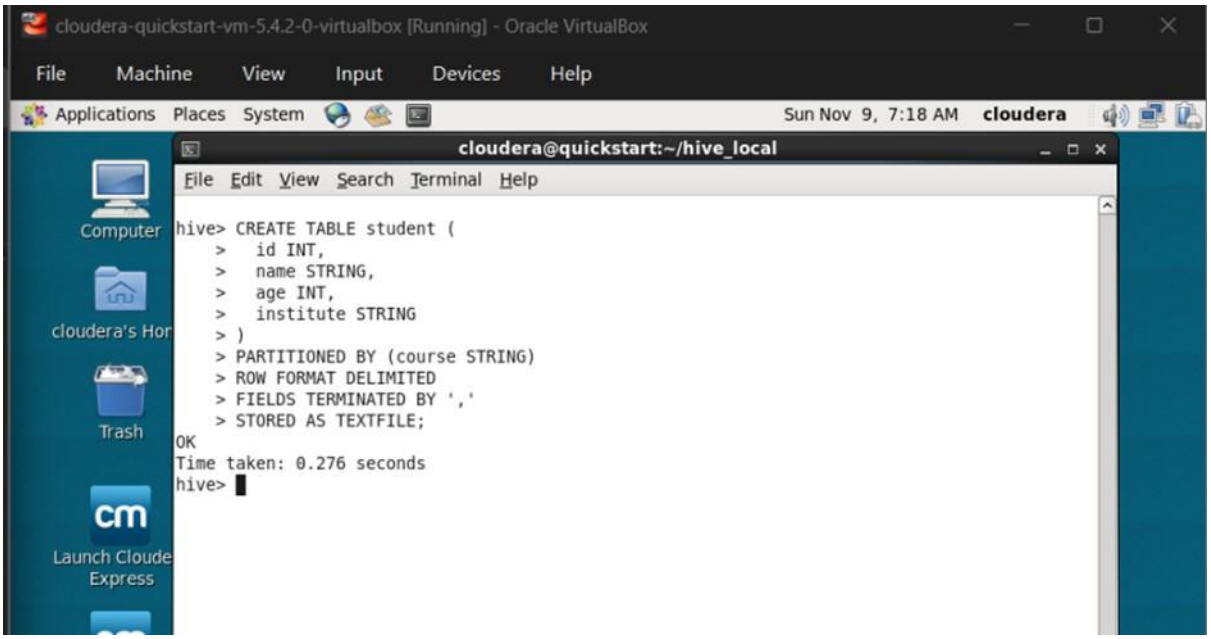
Static Partition

To learn how to store Hive table data in separate directory partitions (for faster queries and organization)

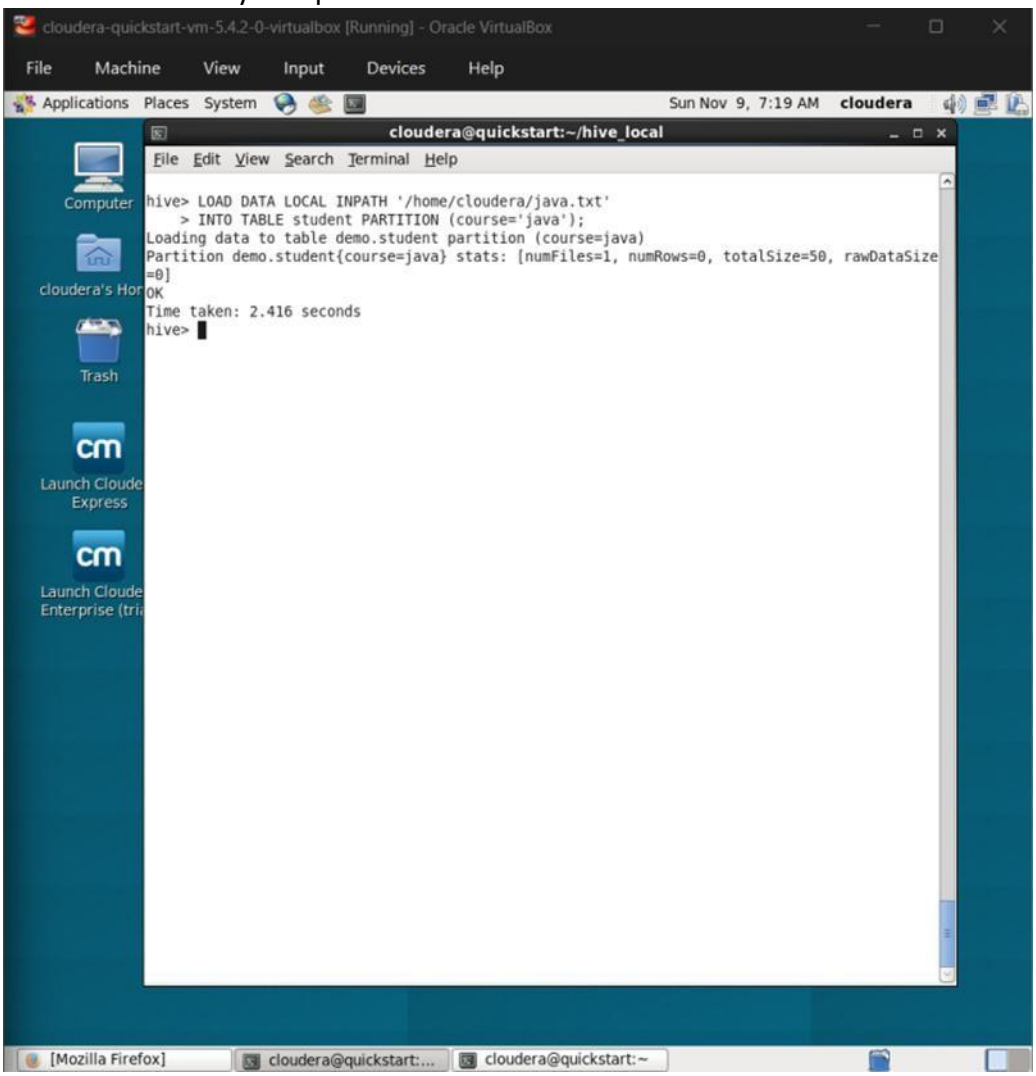


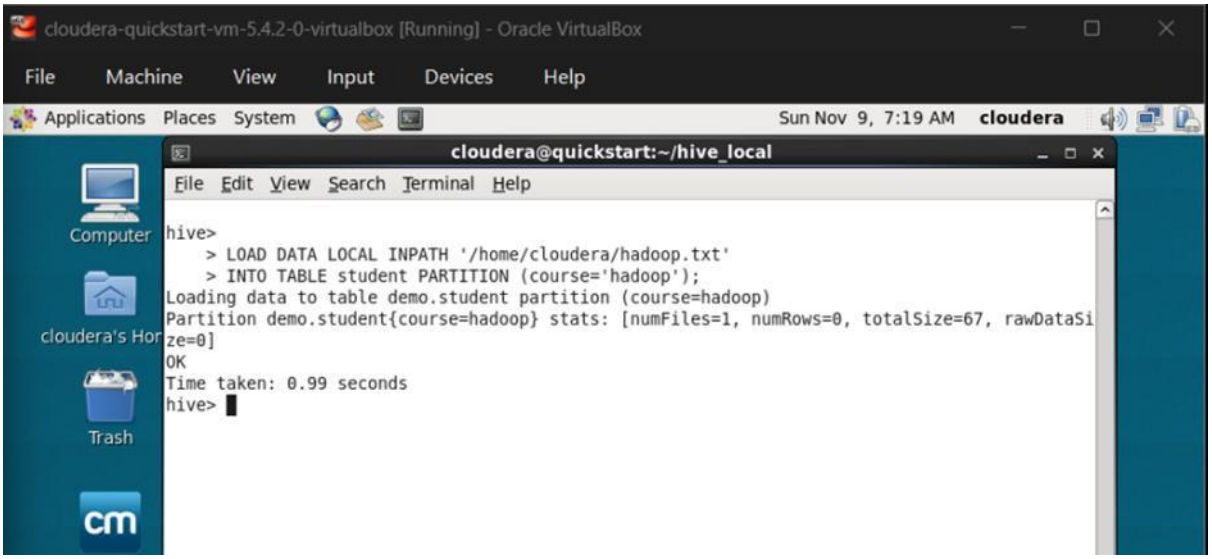
creating table student

BIGDATA

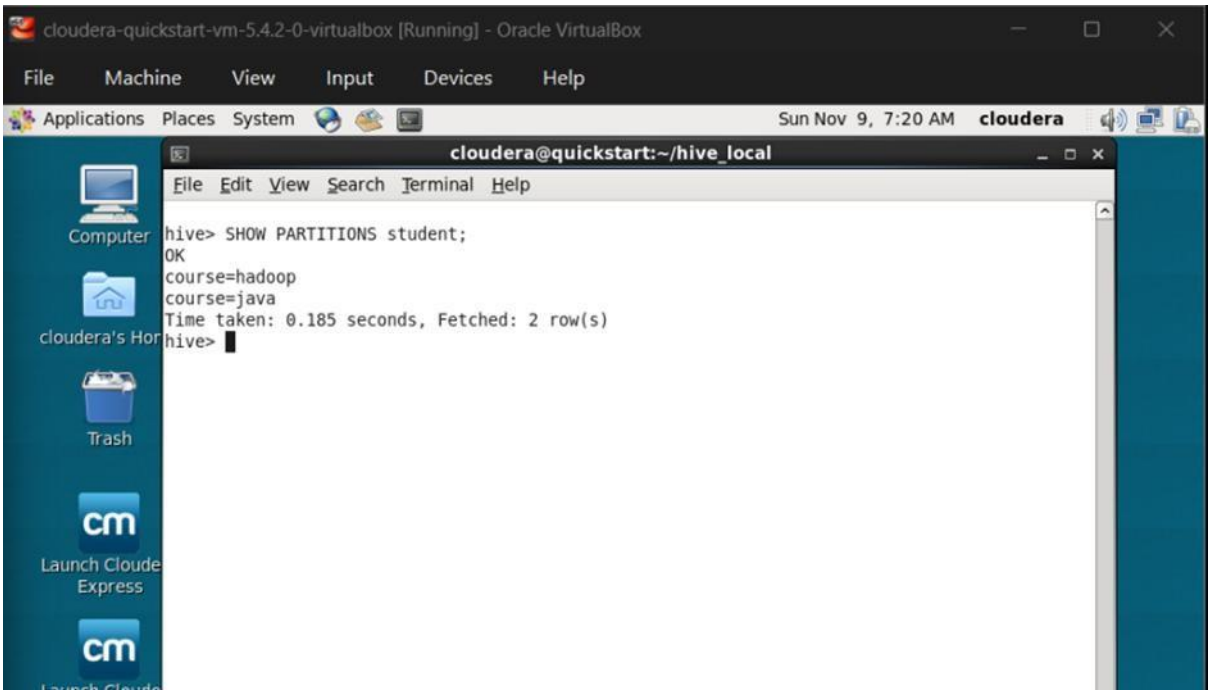


Load data statically into partitions





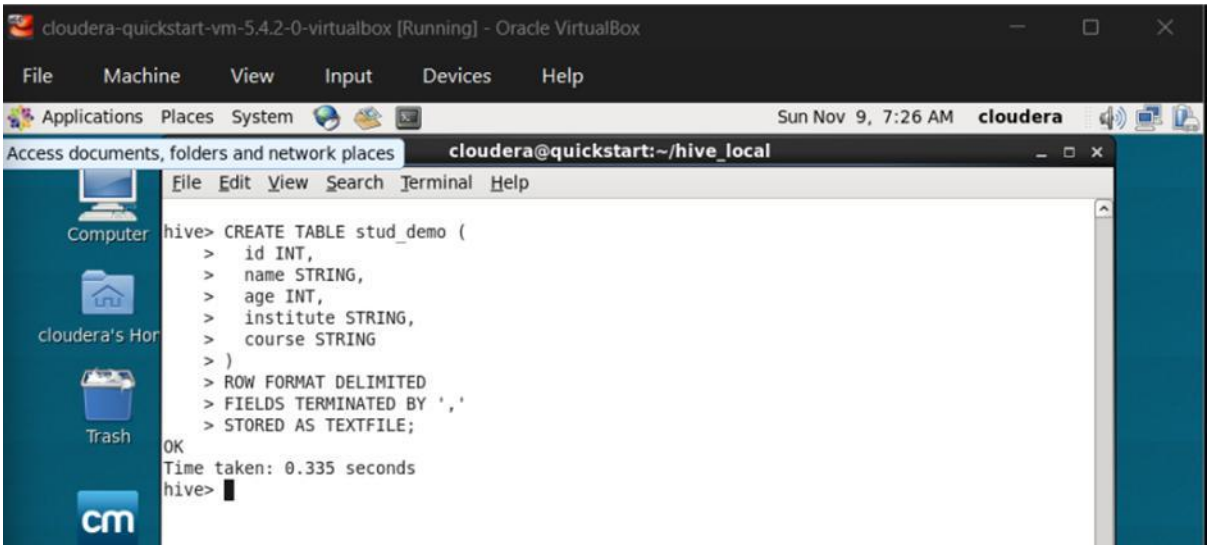
Verify partition structure



Dynamic Partitioning

To automatically create Hive table partitions while inserting data — instead of specifying each partition value manually.

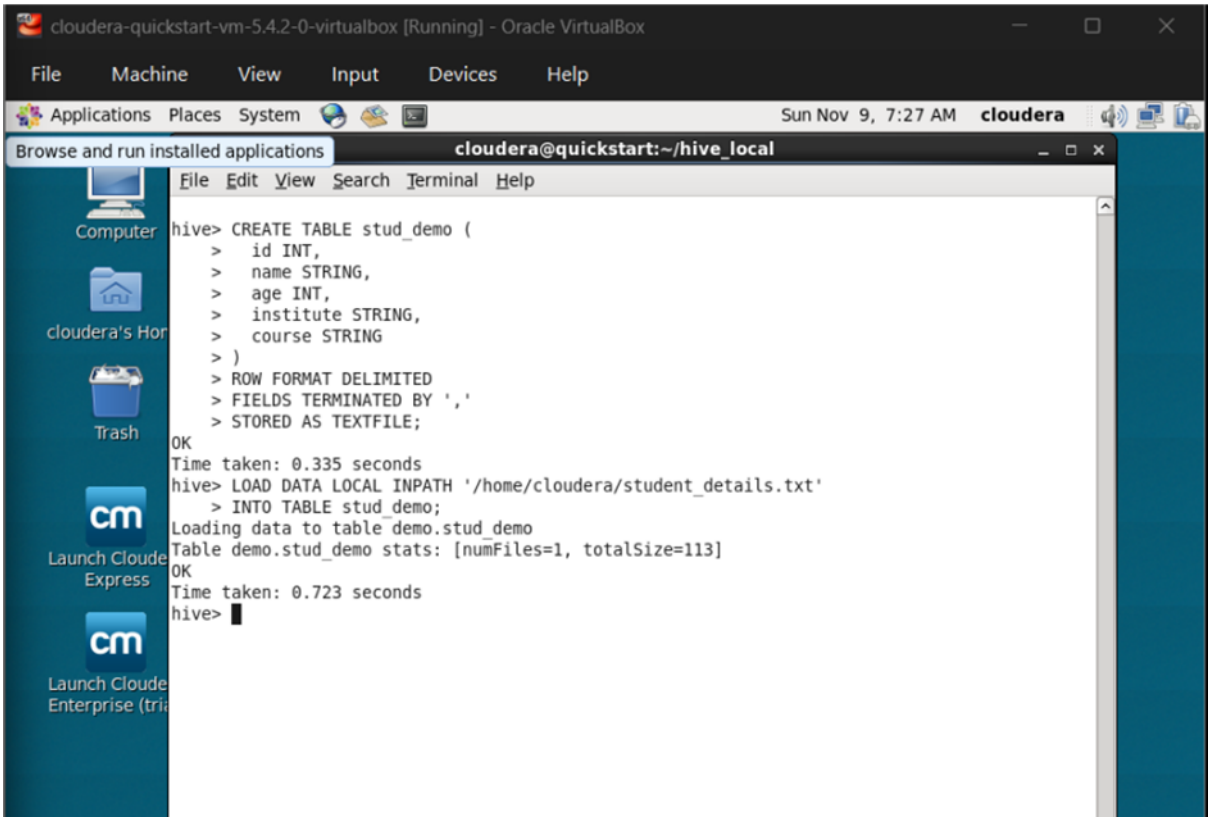
Create the Staging Table

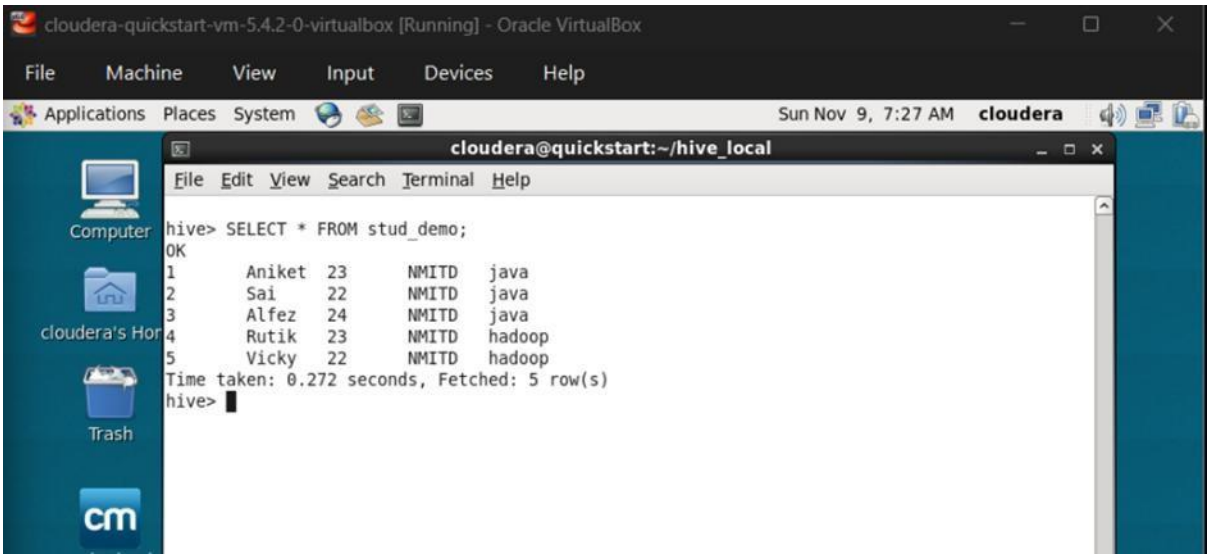


Prepare Combined Input File

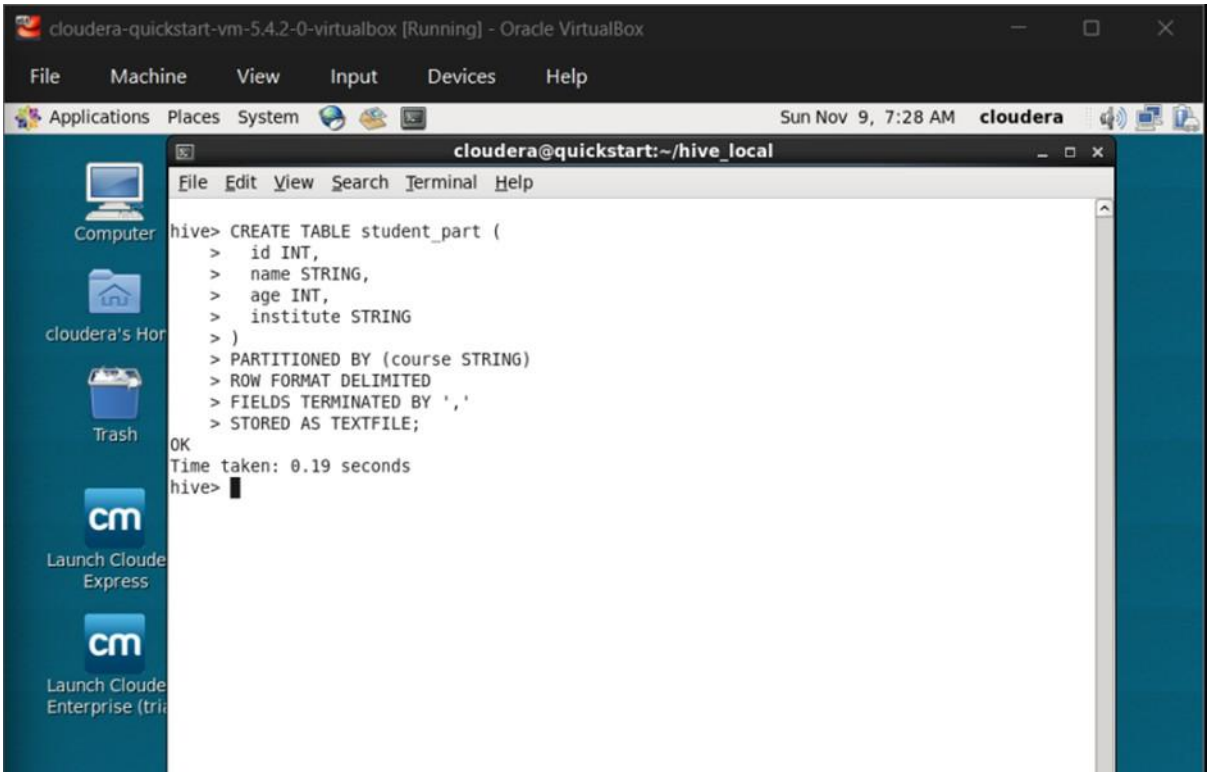
In your Linux terminal, create one combined file that includes both course values

Load Data into the Staging Table



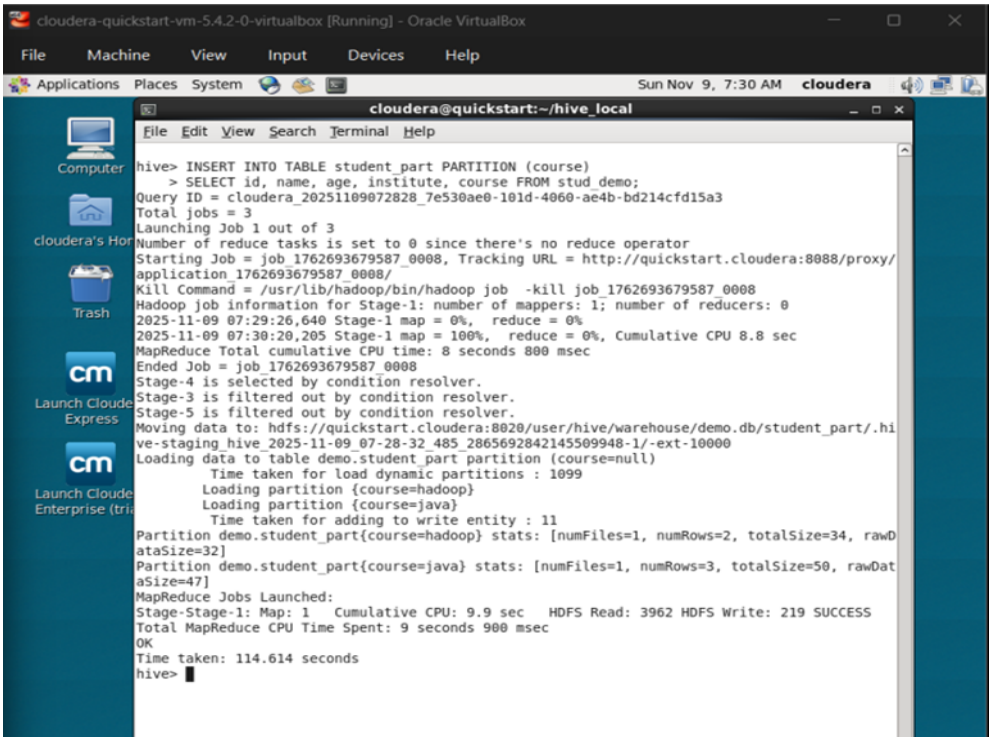


Create the Partitioned Target Table

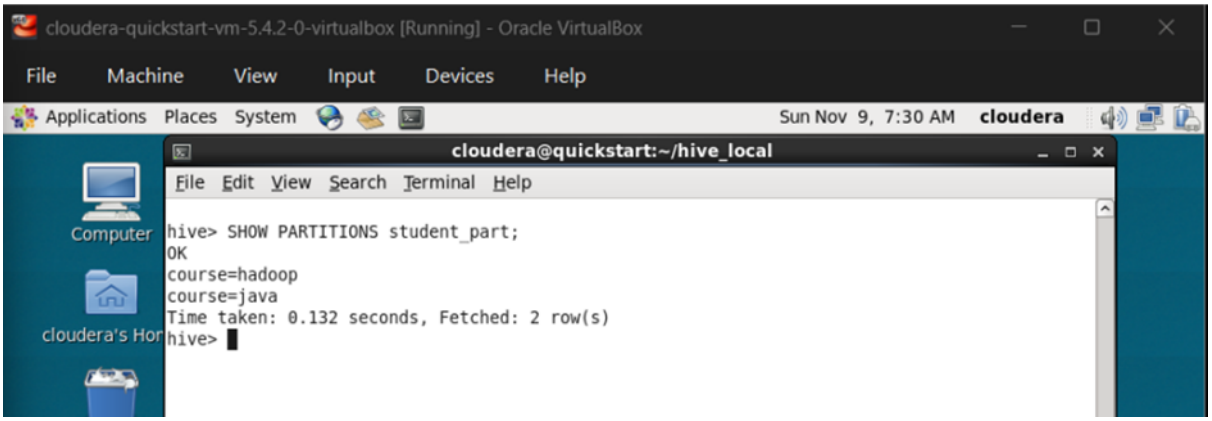


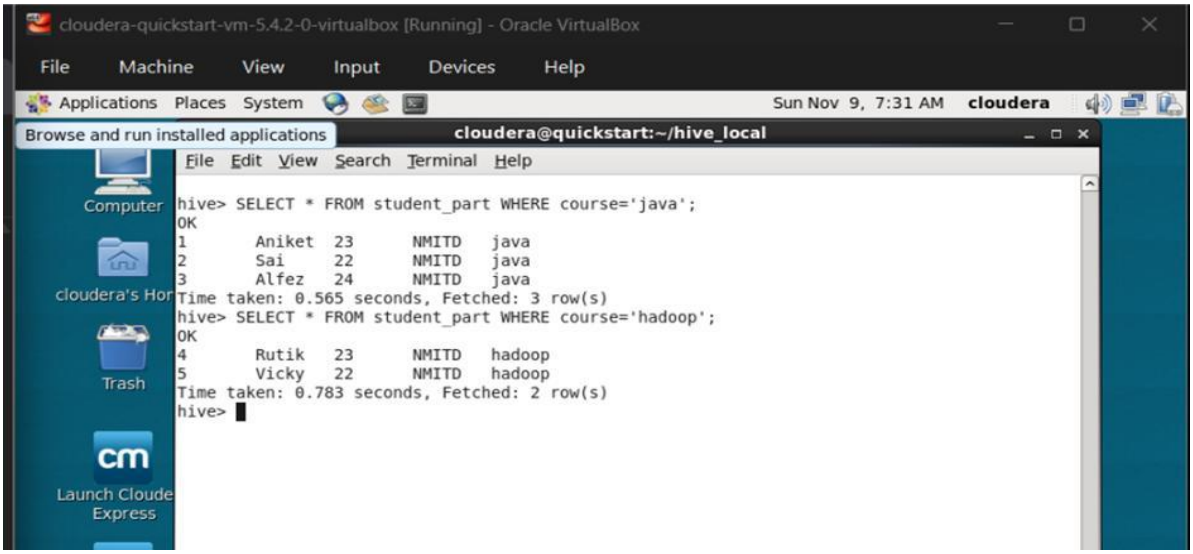
Insert Data Dynamically

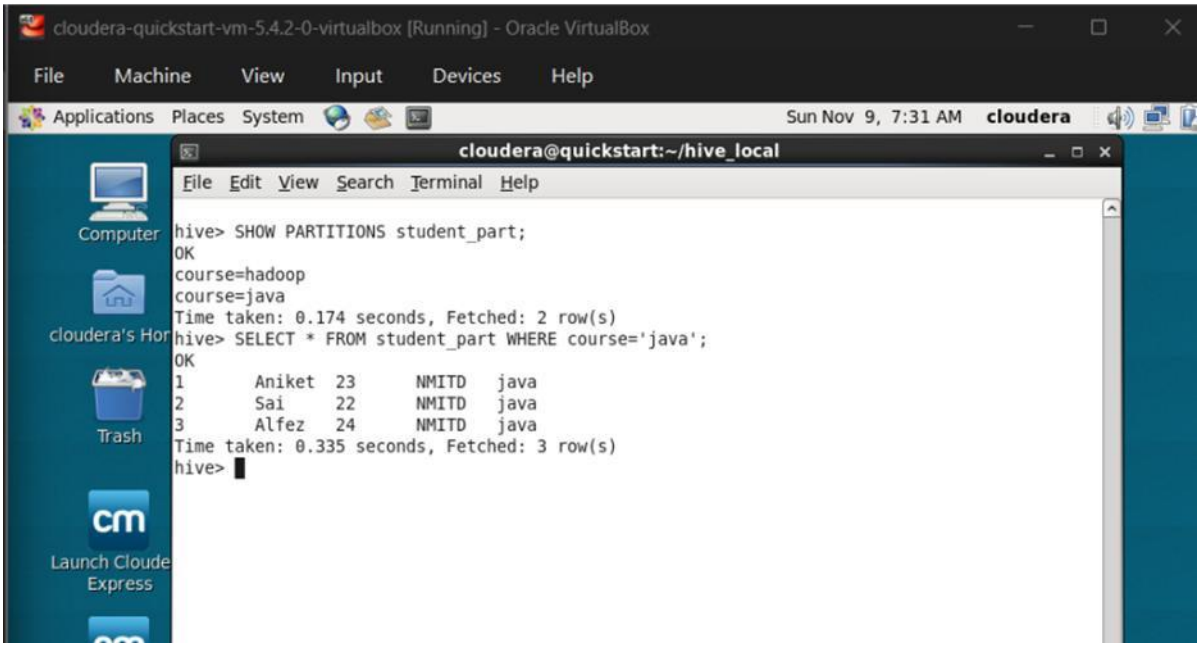
BIGDATA



Verify Dynamic Partitions



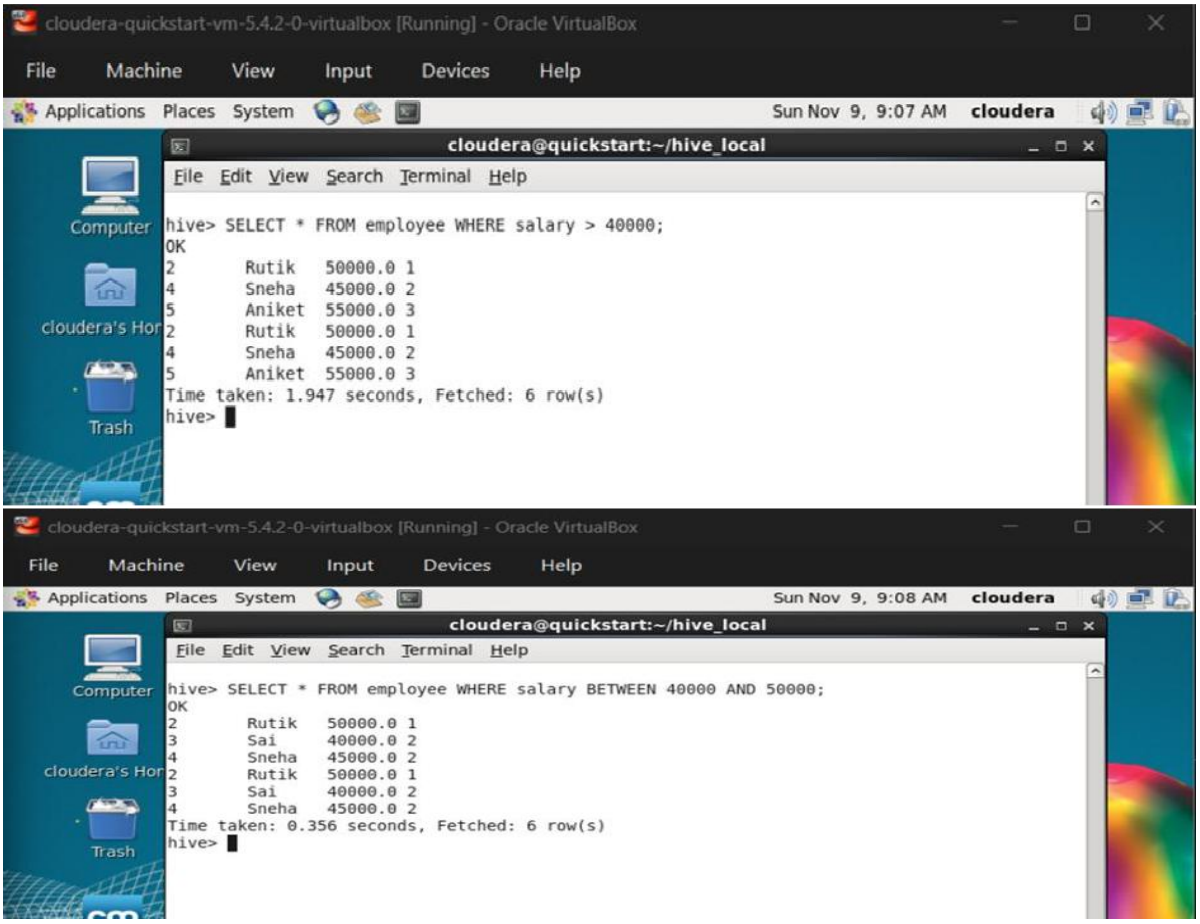




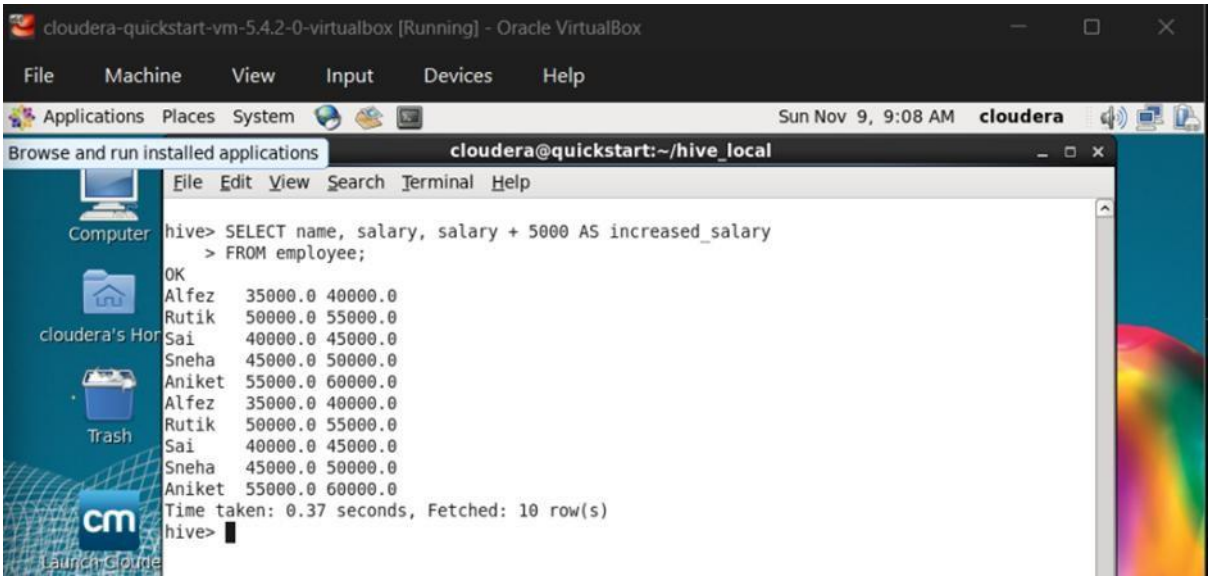
Hive Built-In Operators

To demonstrate the use of Hive's built-in operators for comparison, arithmetic, and logical operations on table data.

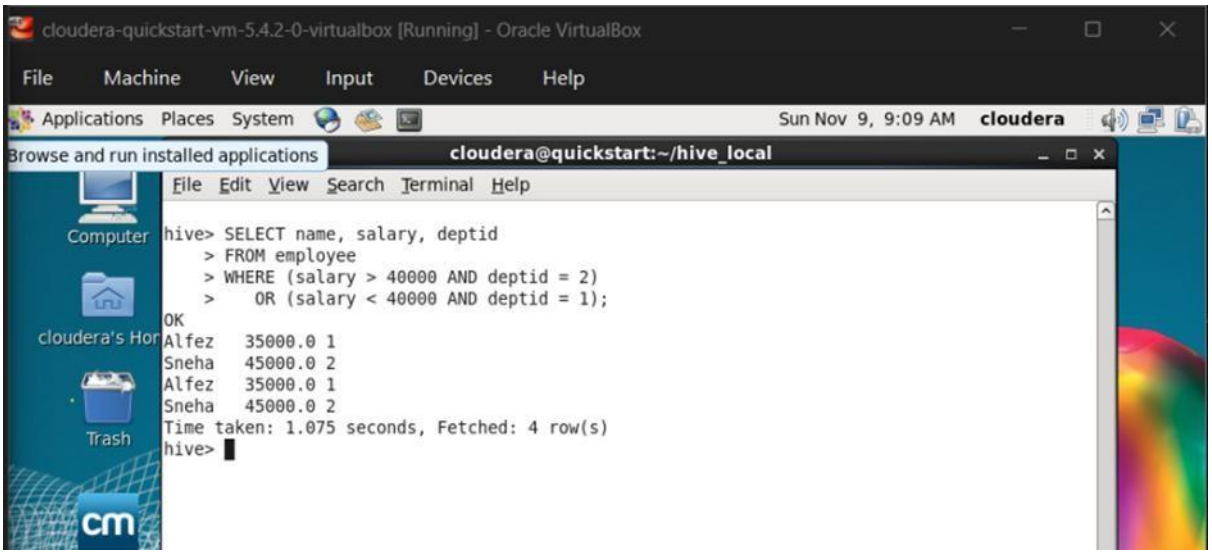
Comparison Operators



Arithmetic Operators

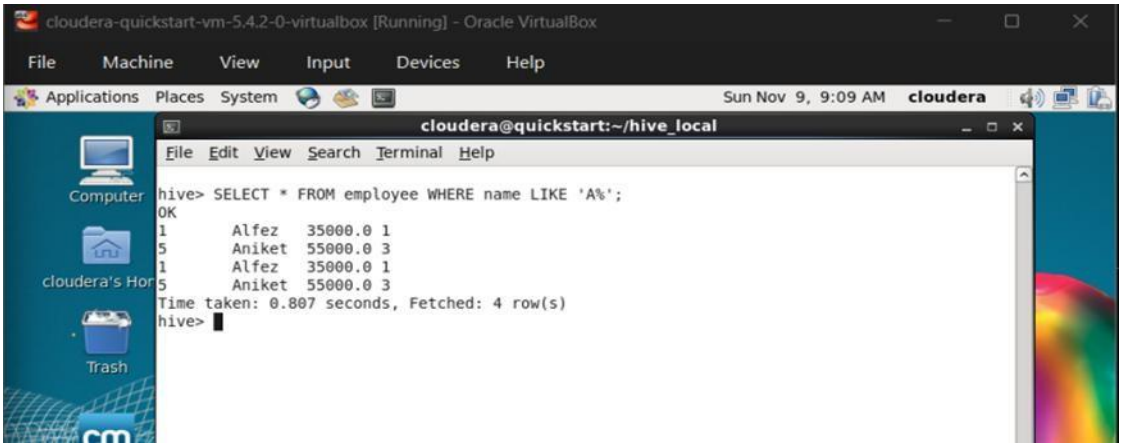


Logical Operators

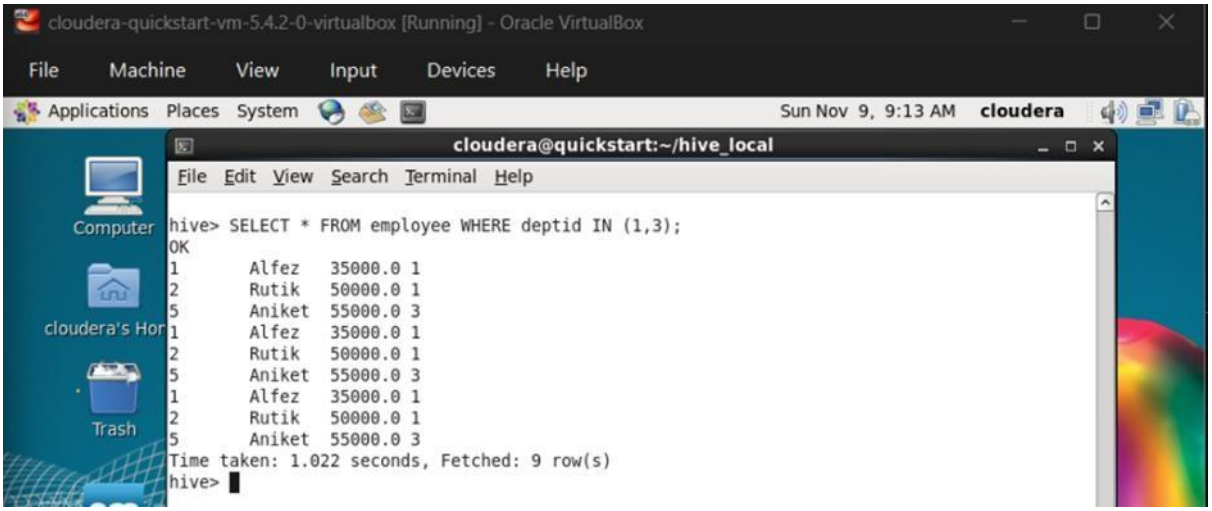


Special Operator Examples

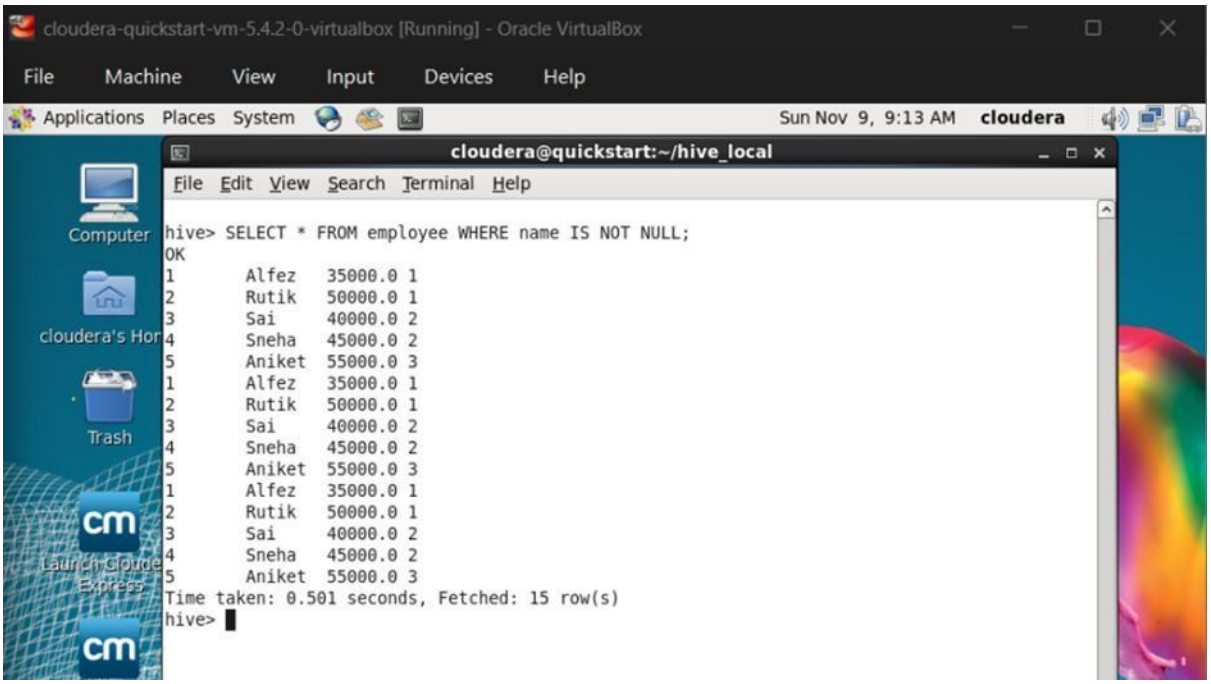
LIKE — pattern matching



IN / NOT IN



IS NULL / IS NOT NULL



Hive Built-In Functions

To demonstrate the use of Hive built-in functions — both aggregate functions (like MAX, MIN, AVG) and string manipulation functions (like UPPER, LOWER, LENGTH, SUBSTR).

Aggregate Functions

SELECT MAX(salary) AS max_salary FROM employee;

BIGDATA

```
cloudera@quickstart:~/hive_local
File  Edit  View  Search  Terminal  Help
hive> SELECT MAX(salary) AS max_salary FROM employee;
Query ID = cloudera_20251109075151_4d528fa2-4040-4a1f-9164-6c670aa6a3d5
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_1762693679587_0013, Tracking URL = http://quickstart.cloudera:8088/proxy/
application_1762693679587_0013/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1762693679587_0013
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2025-11-09 07:52:11,209 Stage-1 map = 0%,  reduce = 0%
2025-11-09 07:52:48,503 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.76 sec
2025-11-09 07:53:25,833 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 12.2 sec
2025-11-09 07:53:30,725 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 17.58 sec
MapReduce Total cumulative CPU time: 17 seconds 580 msec
Ended Job = job_1762693679587_0013
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 17.58 sec   HDFS Read: 6742 HDFS Write: 8
 SUCCESS
Total MapReduce CPU Time Spent: 17 seconds 580 msec
OK
55000.0
Time taken: 124.335 seconds, Fetched: 1 row(s)
```

SELECT MIN(salary) AS min_salary FROM employee

```
hive> SELECT MIN(salary) AS min_salary FROM employee;
Query ID = cloudera_20251109075353_ef1b346e-e3f0-4dba-a99f-4271cc0c992a
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_1762693679587_0014, Tracking URL = http://quickstart.cloudera:8088/proxy/
application_1762693679587_0014/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1762693679587_0014
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2025-11-09 07:54:16,118 Stage-1 map = 0%,  reduce = 0%
```

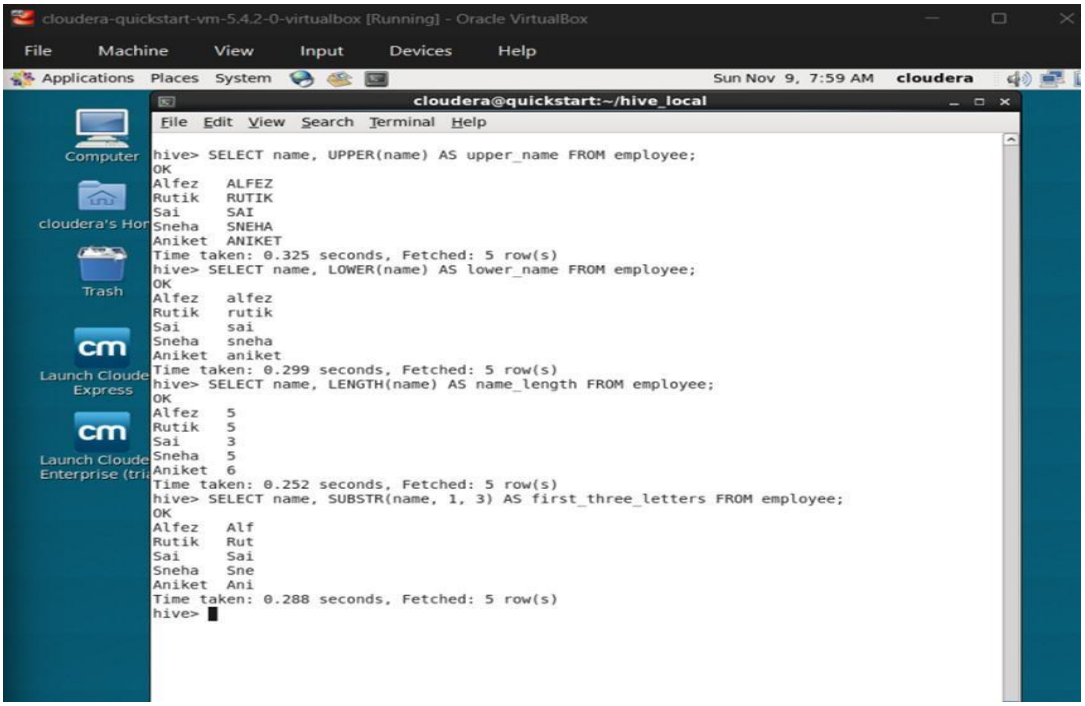
SELECT AVG(salary)AS avg_salary FROM employee

```
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 18.62 sec   HDFS Read: 6749 HDFS Write: 8
 SUCCESS
Total MapReduce CPU Time Spent: 18 seconds 620 msec
OK
35000.0
Time taken: 116.092 seconds, Fetched: 1 row(s)
hive> SELECT AVG(salary) AS avg_salary FROM employee;
Query ID = cloudera_20251109075555_c907aad7-f1b2-408a-9429-efbd991e2eb4
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_1762693679587_0015, Tracking URL = http://quickstart.cloudera:8088/proxy/
application_1762693679587_0015/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1762693679587_0015
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 1
2025-11-09 07:56:23,417 Stage-1 map = 0%,  reduce = 0%
2025-11-09 07:57:01,849 Stage-1 map = 100%,  reduce = 0%, Cumulative CPU 7.94 sec
2025-11-09 07:57:33,105 Stage-1 map = 100%,  reduce = 67%, Cumulative CPU 13.68 sec
2025-11-09 07:57:37,676 Stage-1 map = 100%,  reduce = 100%, Cumulative CPU 18.06 sec
MapReduce Total cumulative CPU time: 18 seconds 60 msec
Ended Job = job_1762693679587_0015
MapReduce Jobs Launched:
Stage-Stage-1: Map: 1  Reduce: 1   Cumulative CPU: 18.06 sec   HDFS Read: 7116 HDFS Write: 8
 SUCCESS
Total MapReduce CPU Time Spent: 18 seconds 60 msec
OK
45000.0
Time taken: 128.181 seconds, Fetched: 1 row(s)
hive> S
```

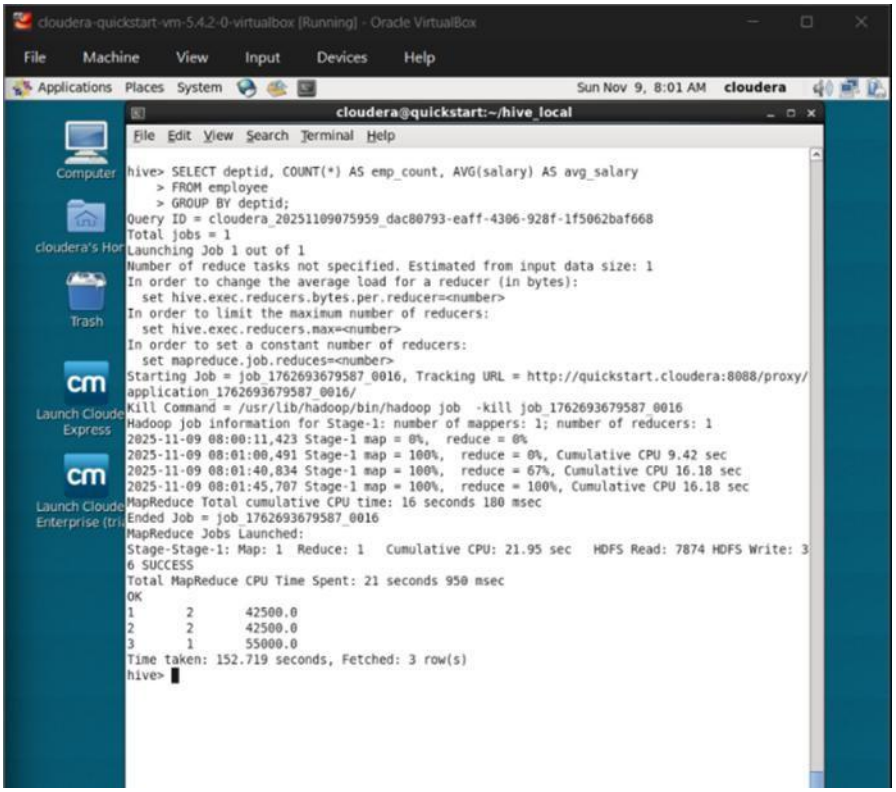
String Functions

Hive provides many string manipulation utilities.

Let’s use your existing employee table to test them:



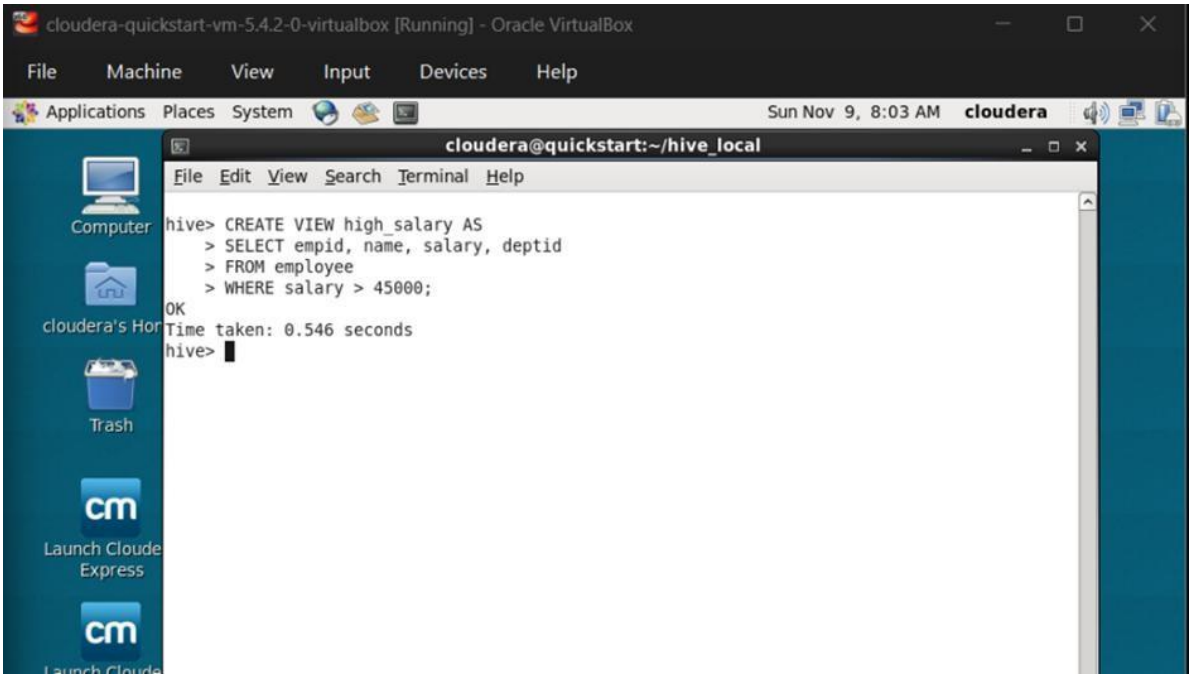
Combining Aggregate & Group



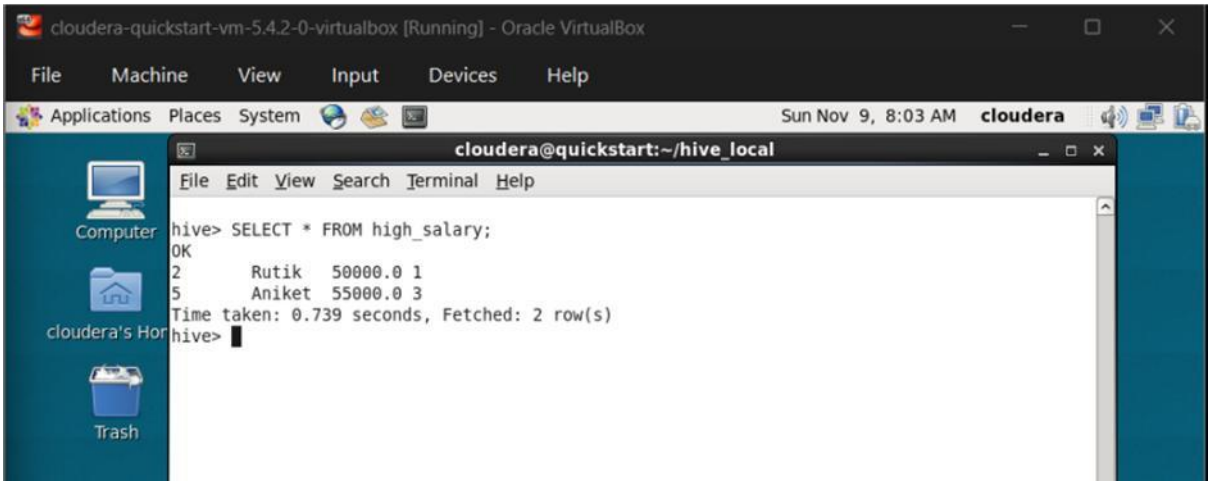
Hive Views

Create and query views (virtual tables) & Understand indexes in Hive (used for query optimization, optional based on VM support).

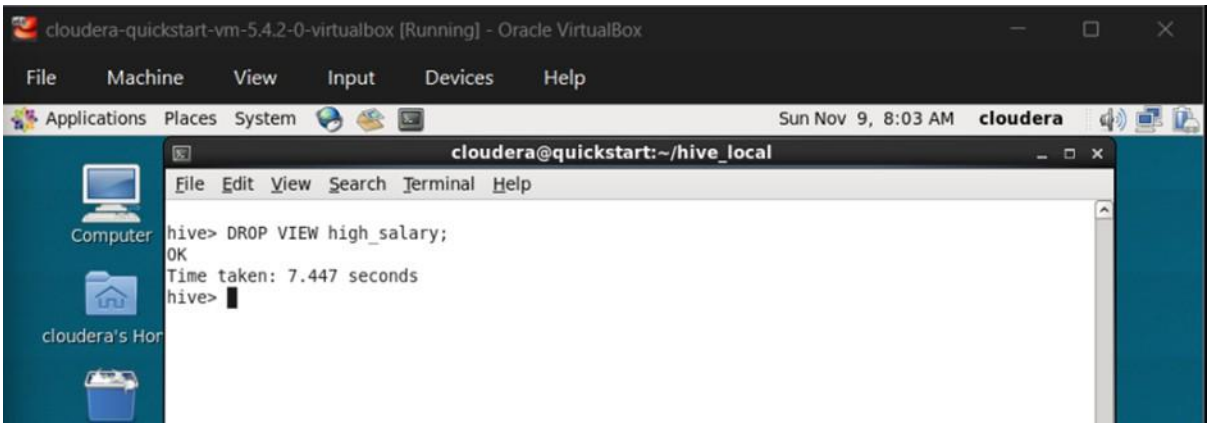
Create a View



Query the view



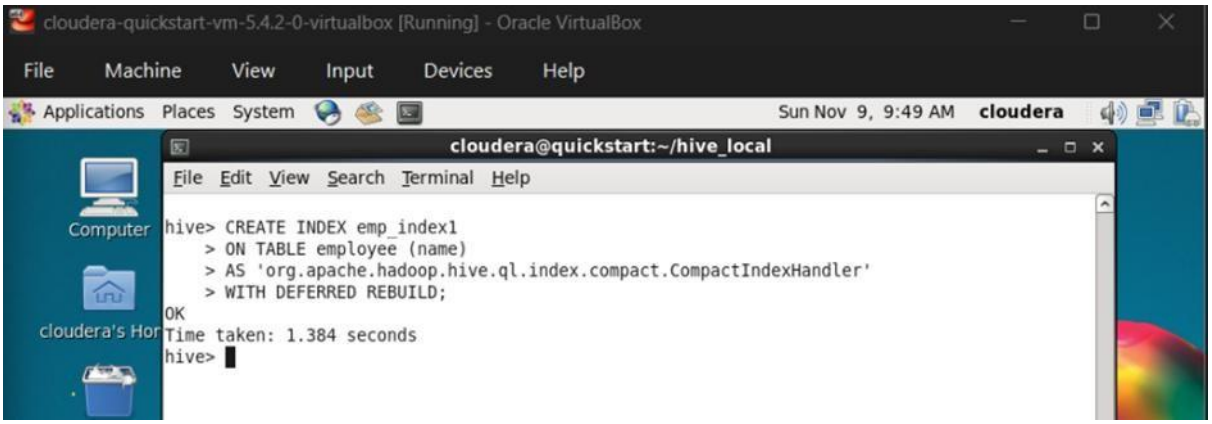
Drop the View



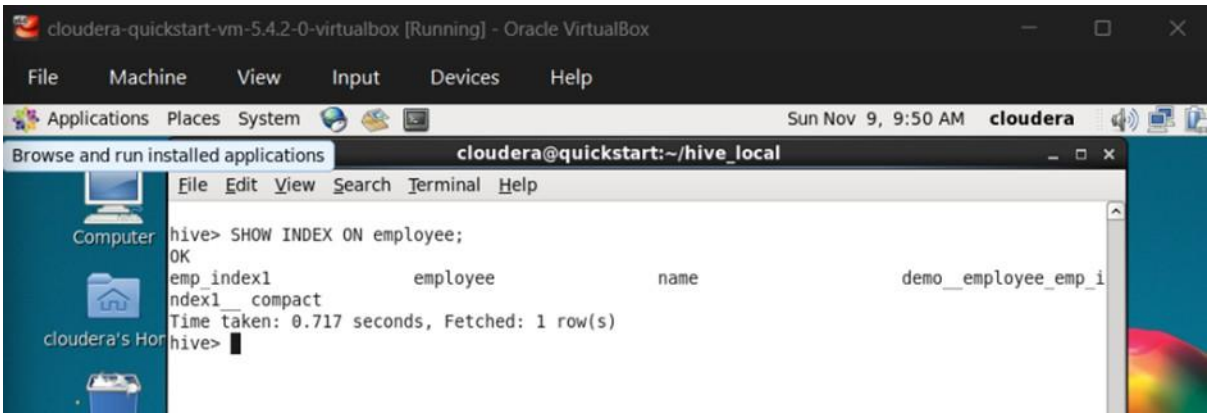
INDEXES

Create an Index - This command creates a metadata index on the name column — intended to speed up queries that filter by name.

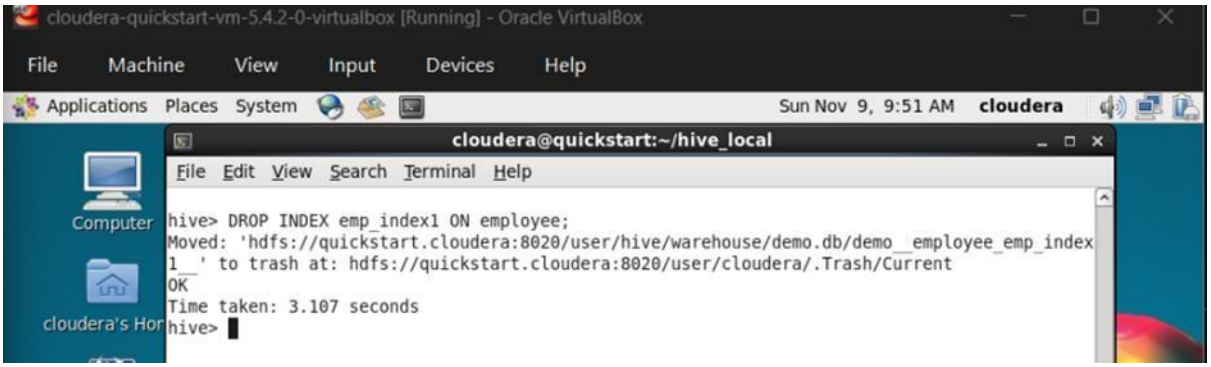
The WITH DEFERRED REBUILD part tells Hive to create the index structure later.



Show Existing Indexes



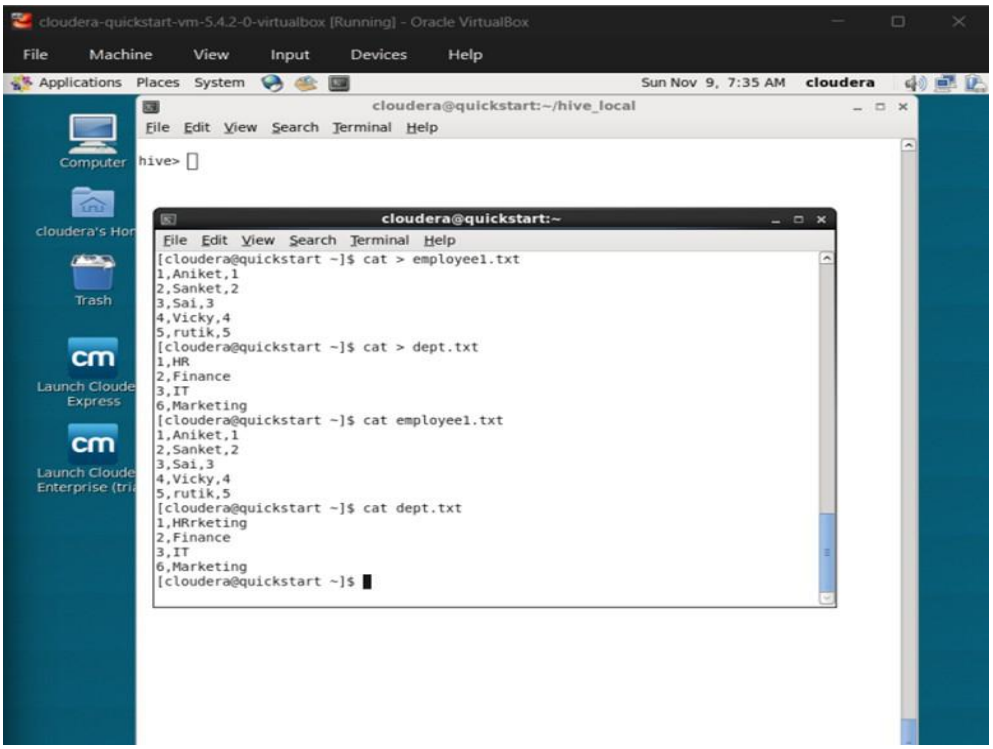
Drop the Index



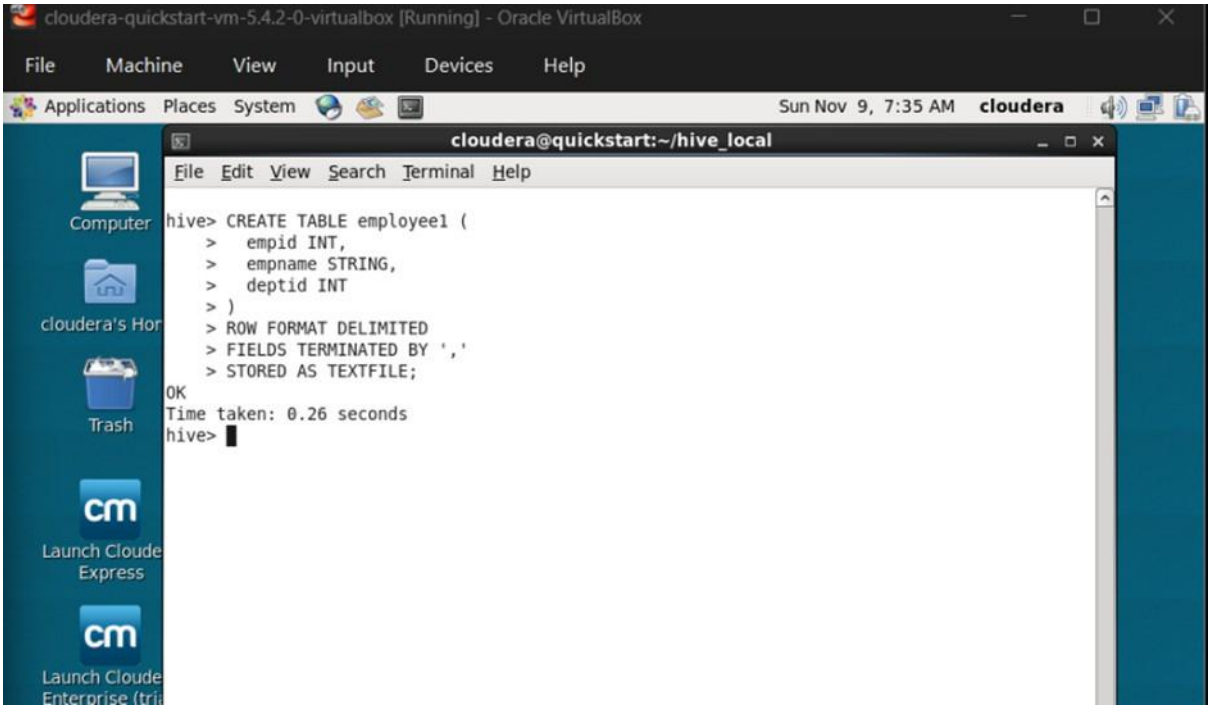
HiveQL: Select Where, Select OrderBy, Select GroupBy, Select Joins

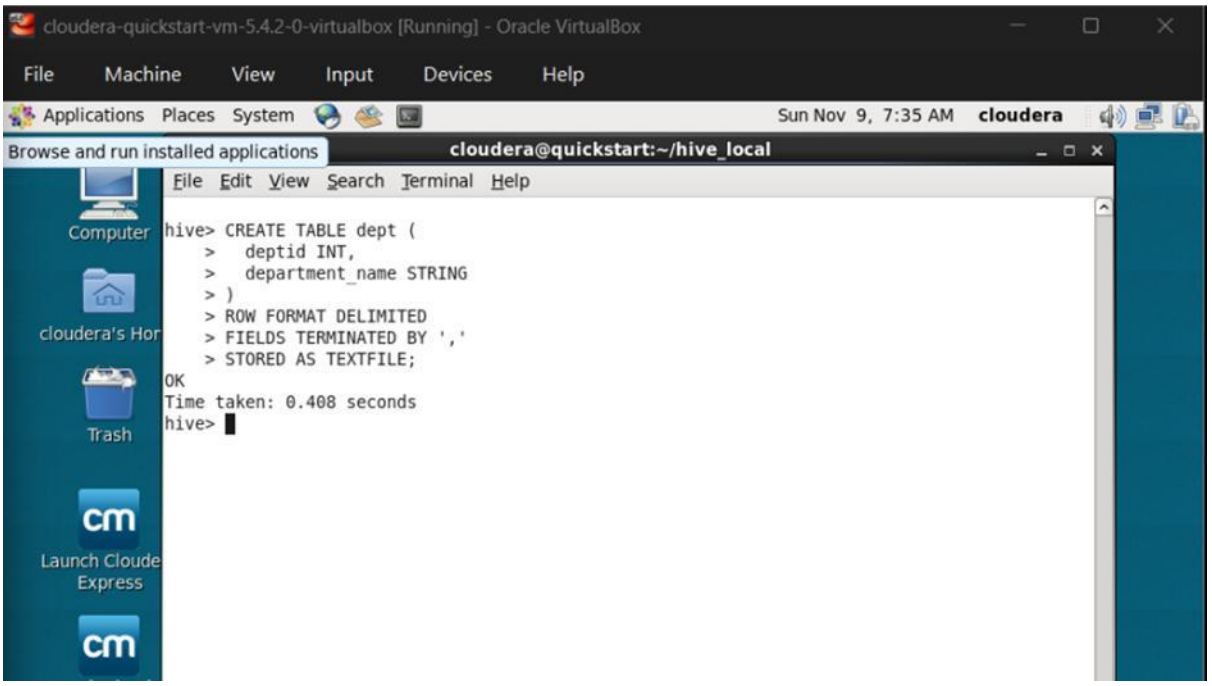
Create Sample Input Files

In your Linux terminal, create two CSV files:

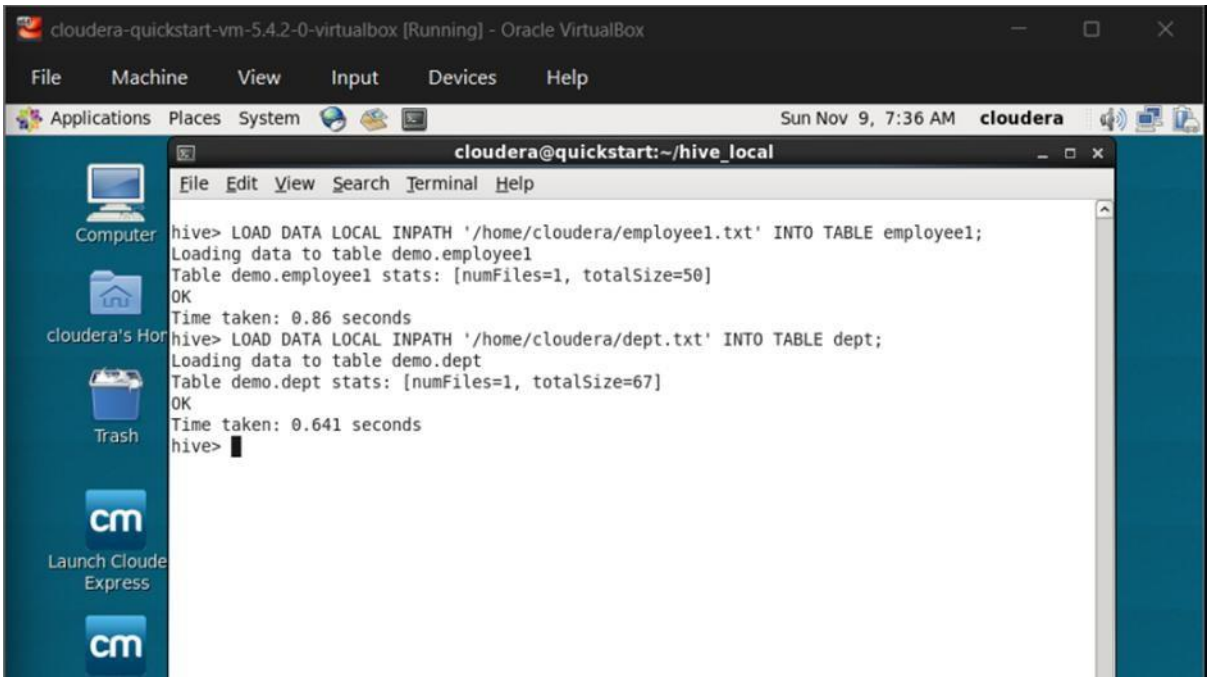


Create the Tables in Hive





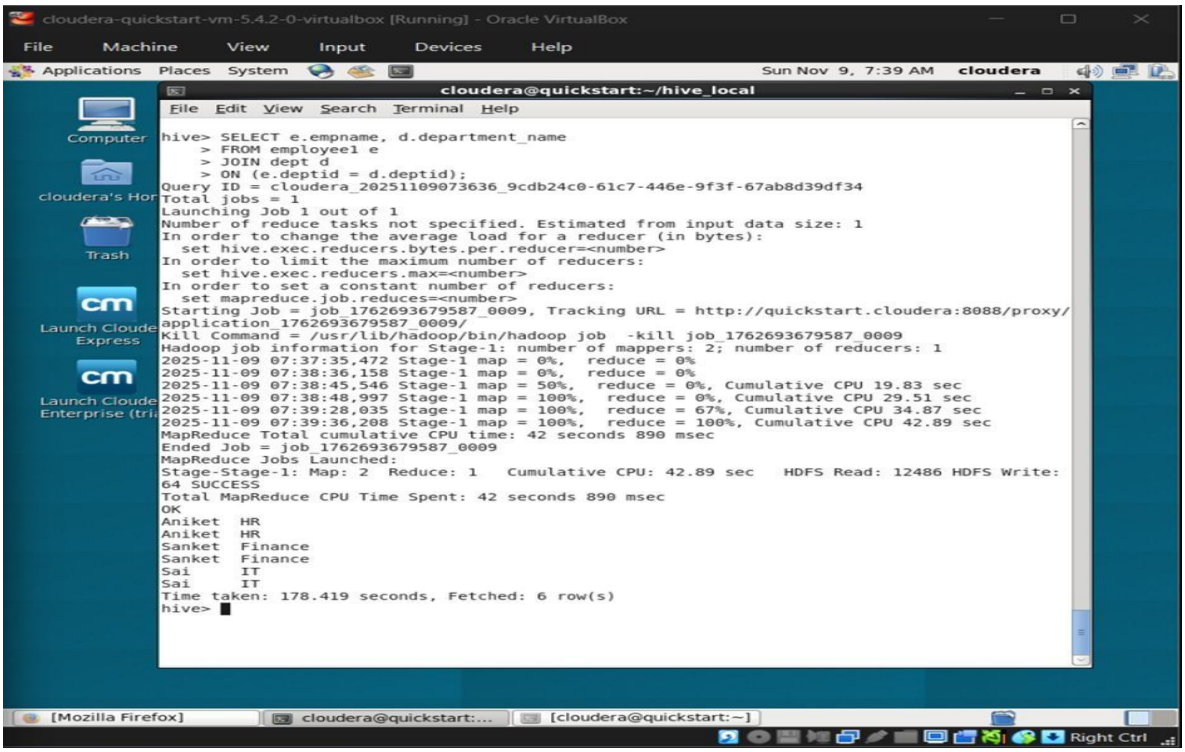
Load Data into the Tables



Disable Auto Join Optimization
This prevents Hive from converting your join into a map-side broadcast join, which can skip reducers for small tables.
Disabling it ensures you see a MapReduce job plan.

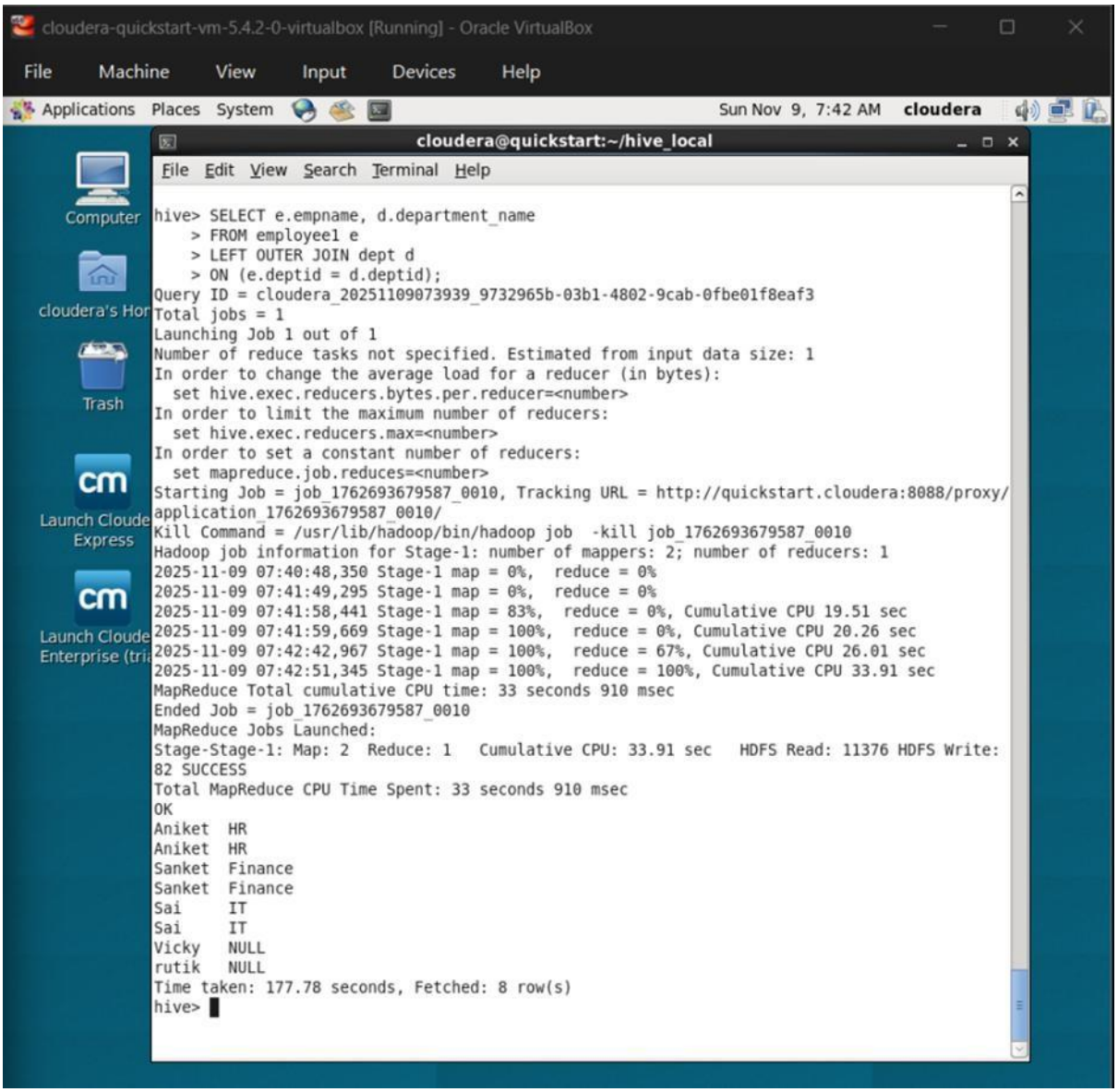
hive.auto.convert.join=false

INNER JOIN



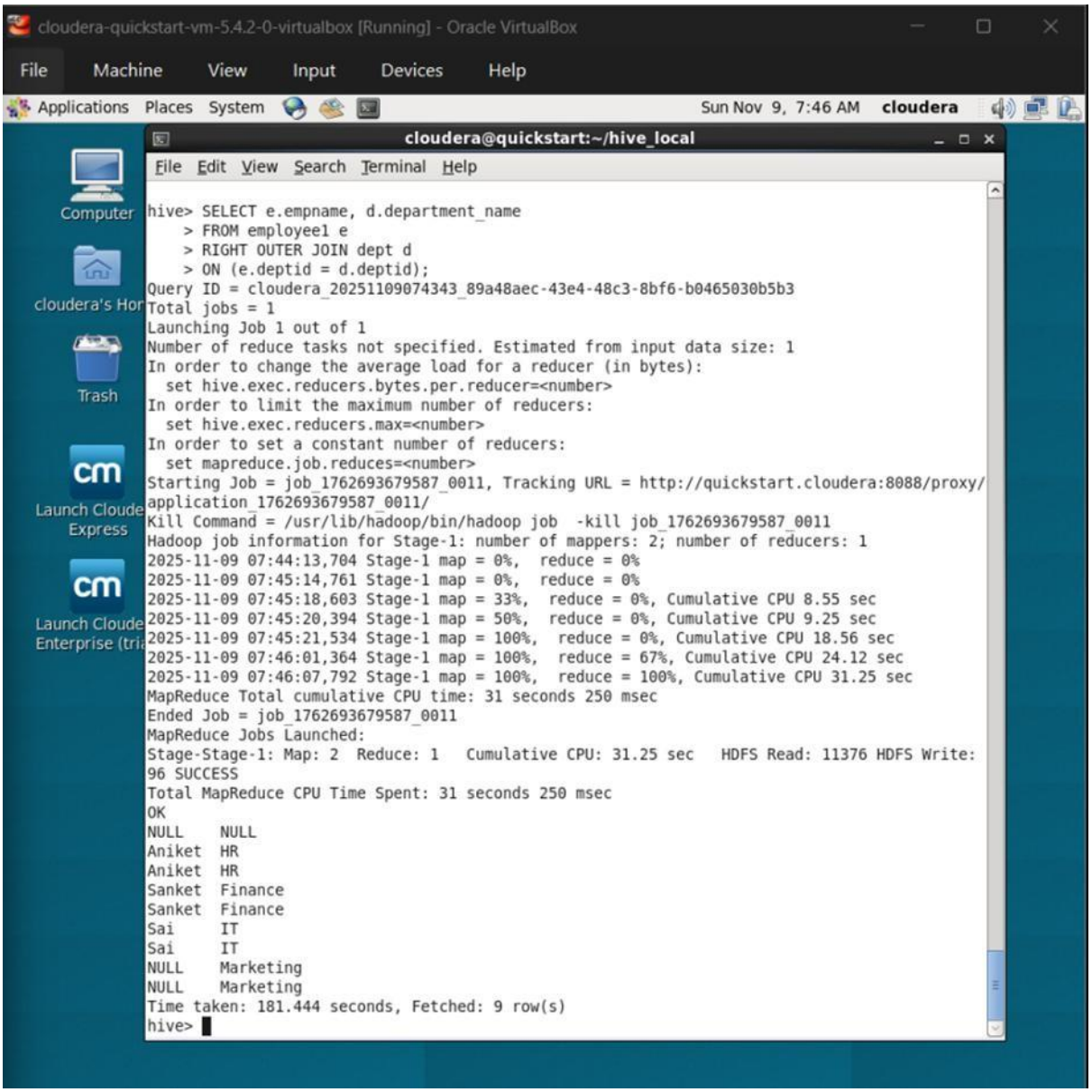
BIGDATA

LEFT OUTER JOIN



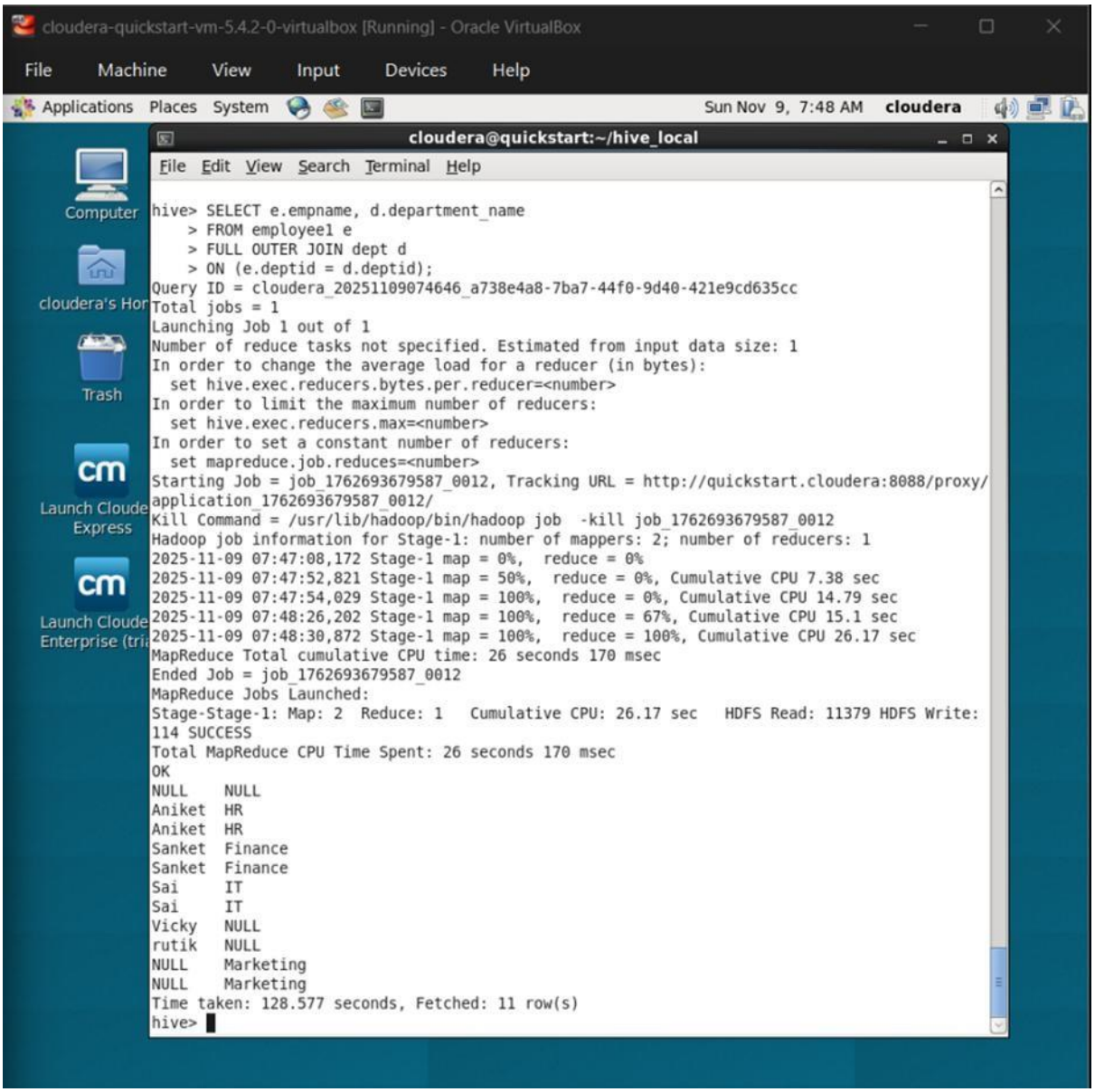
RIGHT OUTER JOIN

BIGDATA

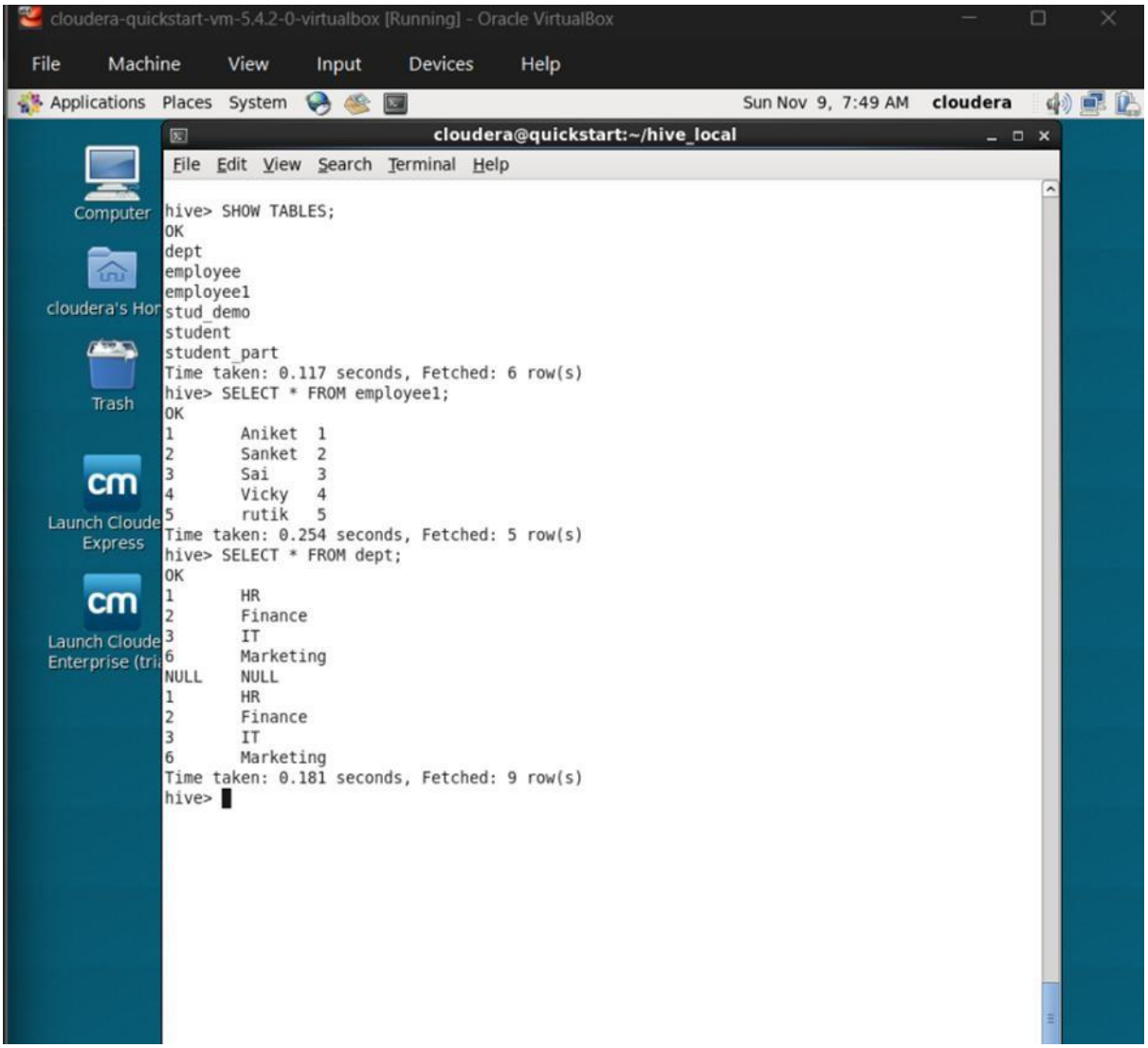


FULL OUTER JOIN

BIGDATA

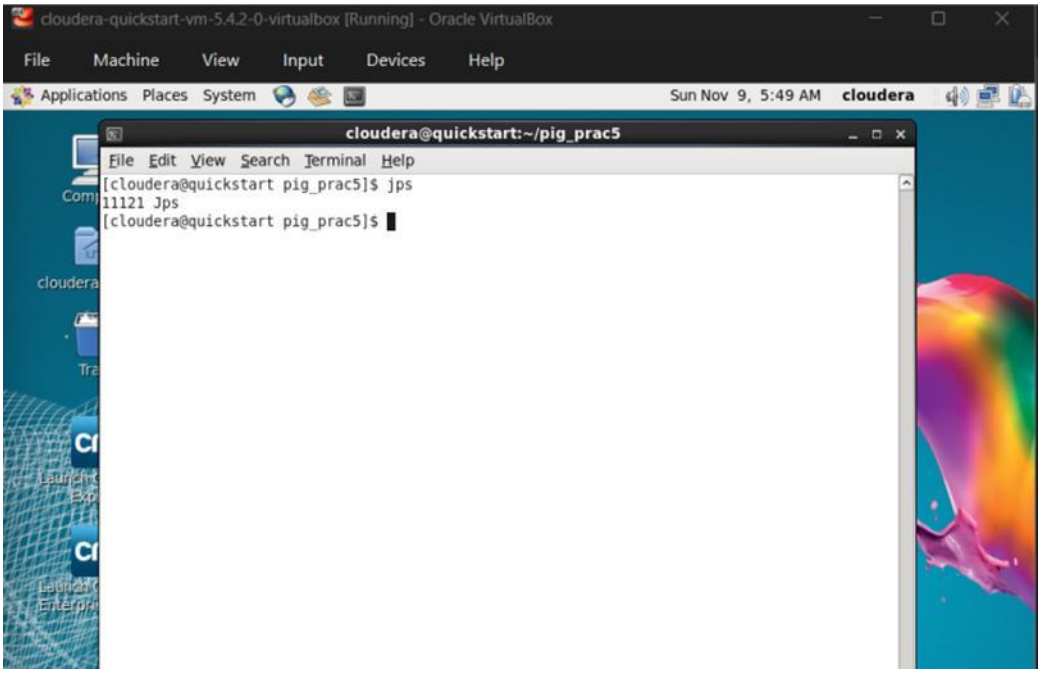

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Nov 9, 7:48 AM cloudera
cloudera@quickstart:~/hive_local
File Edit View Search Terminal Help
Computer
cloudera's Hor
Trash
Launch Cloude Express
Launch Cloude Enterprise (tri
hive> SELECT e.empname, d.department_name
> FROM employee1 e
> FULL OUTER JOIN dept d
> ON (e.deptid = d.deptid);
Query ID = cloudera_20251109074646_a738e4a8-7ba7-44f0-9d40-421e9cd635cc
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_1762693679587_0012, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1762693679587_0012/
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_1762693679587_0012
Hadoop job information for Stage-1: number of mappers: 2; number of reducers: 1
2025-11-09 07:47:08,172 Stage-1 map = 0%, reduce = 0%
2025-11-09 07:47:52,821 Stage-1 map = 50%, reduce = 0%, Cumulative CPU 7.38 sec
2025-11-09 07:47:54,029 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 14.79 sec
2025-11-09 07:48:26,202 Stage-1 map = 100%, reduce = 67%, Cumulative CPU 15.1 sec
2025-11-09 07:48:30,872 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 26.17 sec
MapReduce Total cumulative CPU time: 26 seconds 170 msec
Ended Job = job_1762693679587_0012
MapReduce Jobs Launched:
Stage-Stage-1: Map: 2 Reduce: 1 Cumulative CPU: 26.17 sec HDFS Read: 11379 HDFS Write: 114 SUCCESS
Total MapReduce CPU Time Spent: 26 seconds 170 msec
OK
NULL NULL
Aniket HR
Aniket HR
Sanket Finance
Sanket Finance
Sai IT
Sai IT
Vicky NULL
rutik NULL
NULL Marketing
NULL Marketing
Time taken: 128.577 seconds, Fetched: 11 row(s)
hive>

BIGDATA

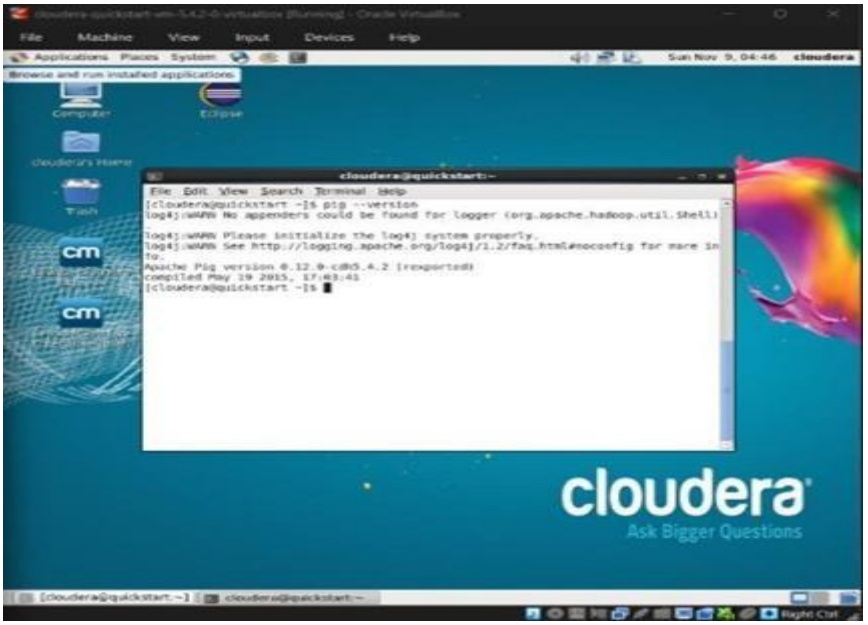

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Nov 9, 7:49 AM cloudera
cloudera@quickstart:~/hive_local
File Edit View Search Terminal Help
Computer
cloudera's Hor
Trash
Launch Cloude Express
Launch Cloude Enterprise (tri
hive> SHOW TABLES;
OK
dept
employee
employee1
stud_demo
student
student_part
Time taken: 0.117 seconds, Fetched: 6 row(s)
hive> SELECT * FROM employee1;
OK
1 Aniket 1
2 Sanket 2
3 Sai 3
4 Vicky 4
5 rutik 5
Time taken: 0.254 seconds, Fetched: 5 row(s)
hive> SELECT * FROM dept;
OK
1 HR
2 Finance
3 IT
6 Marketing
NULL NULL
1 HR
2 Finance
3 IT
6 Marketing
Time taken: 0.181 seconds, Fetched: 9 row(s)
hive>

5. Pig

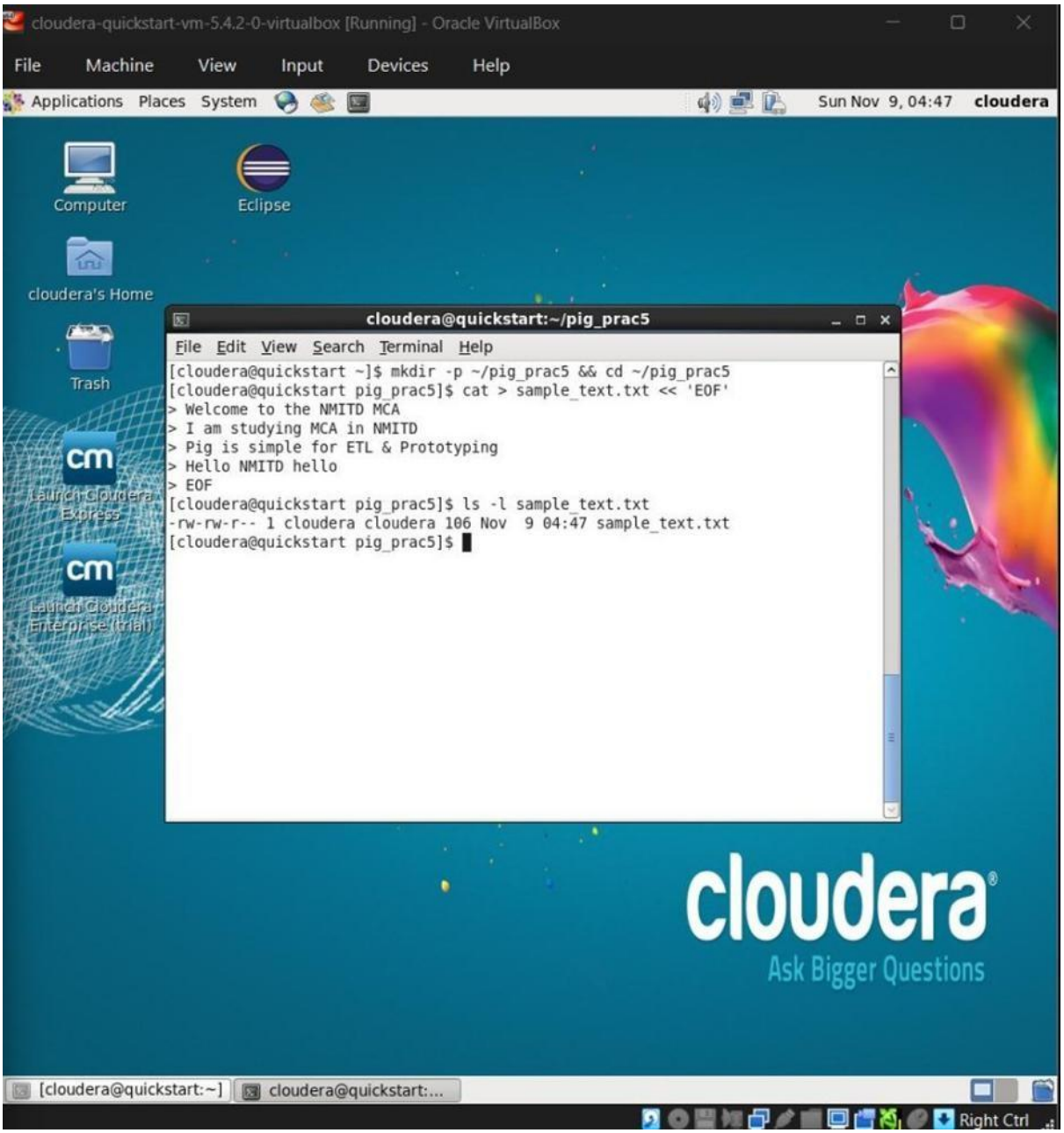
Start the Cloudera Quickstart VM
Launch the VirtualBox and start the Cloudera Quickstart virtual machine.
Confirm that Hadoop daemons (NameNode, DataNode, etc.) are running



Verify Pig Installation

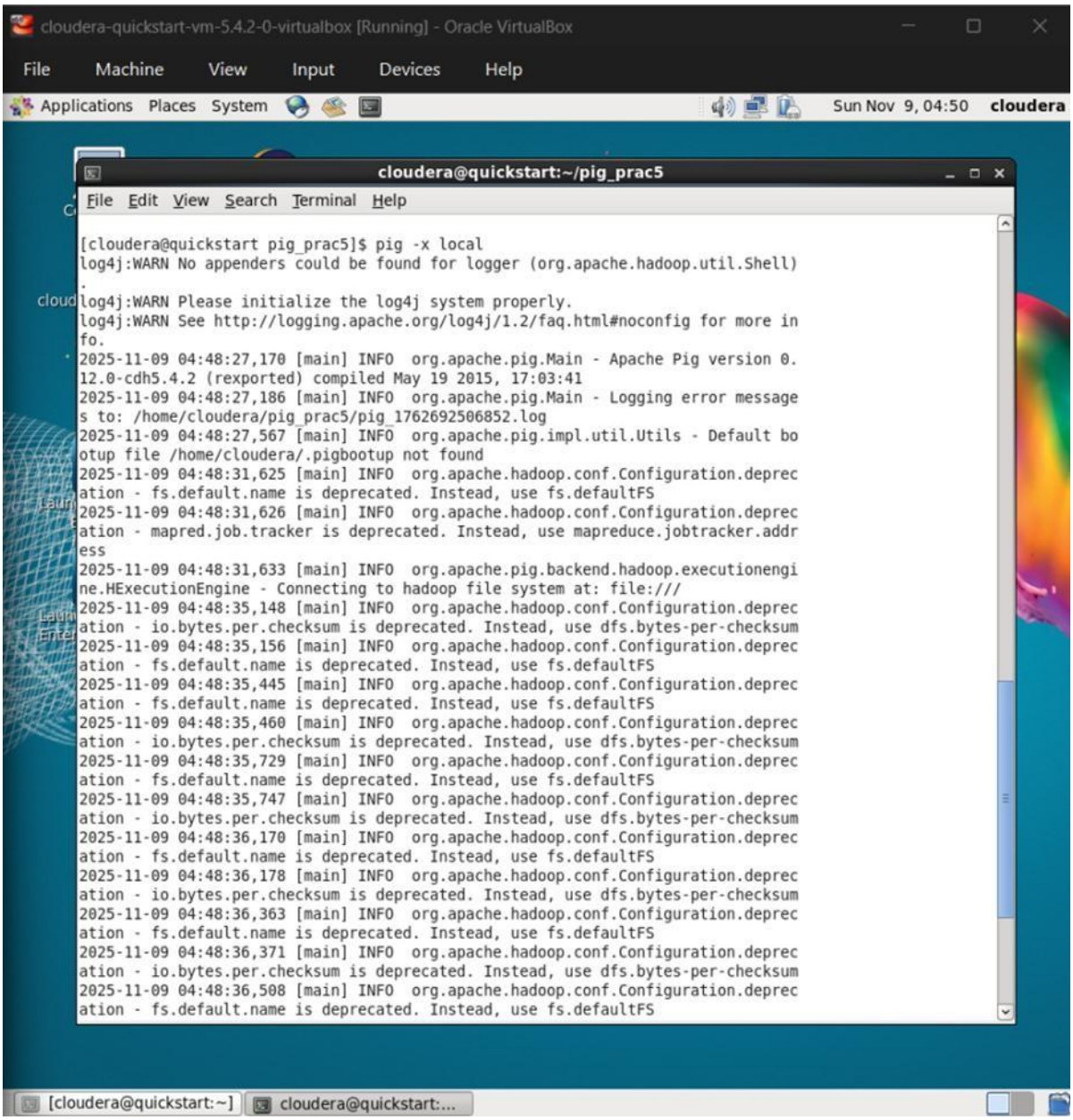


Create Working Directory and Dataset
Create a working directory:
Create a sample text file:
Verify the file

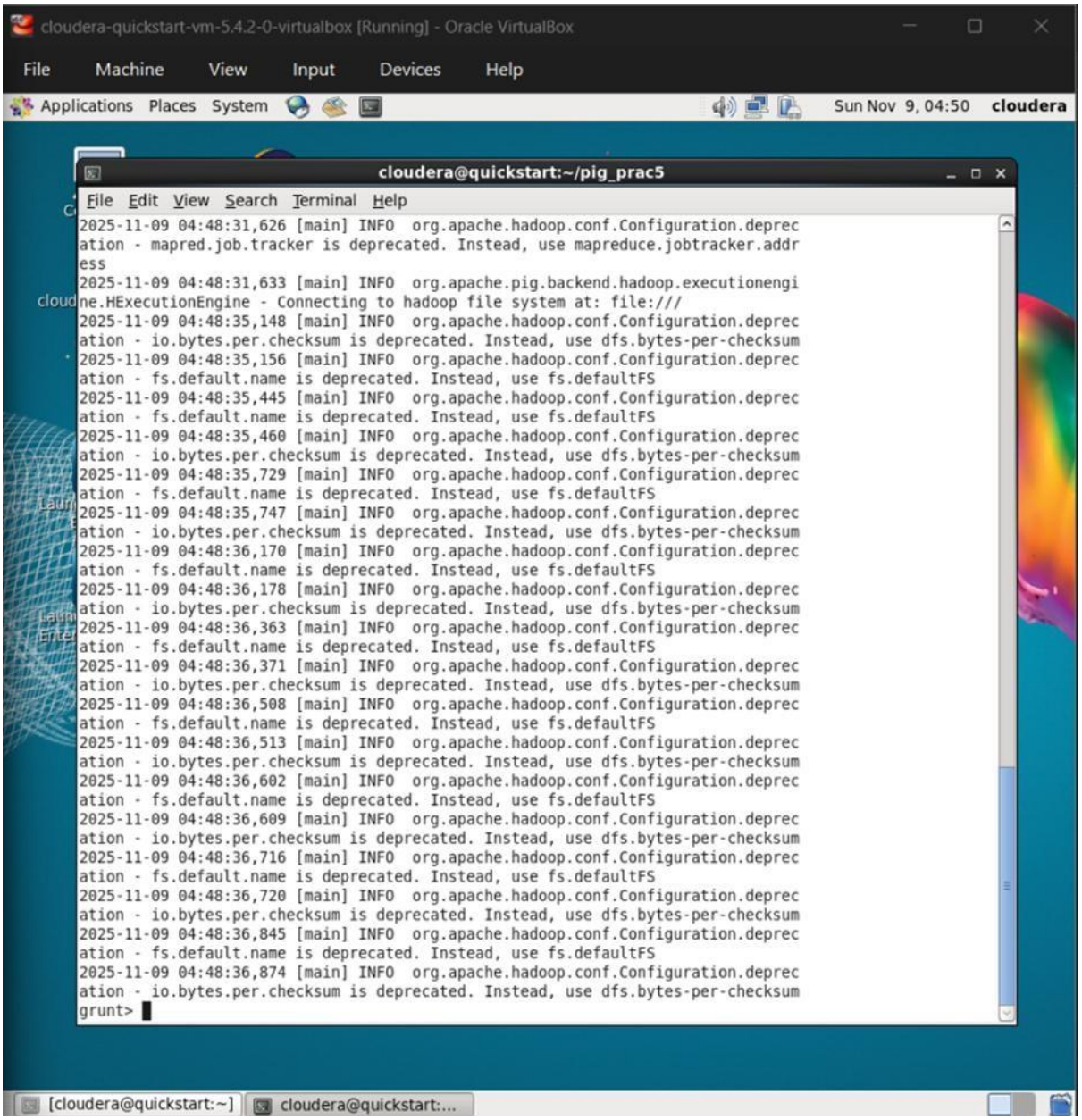


Run Pig in Local Mode
Start Pig shell:
Launch Pig Grunt shell to test basic commands locally

BIGDATA

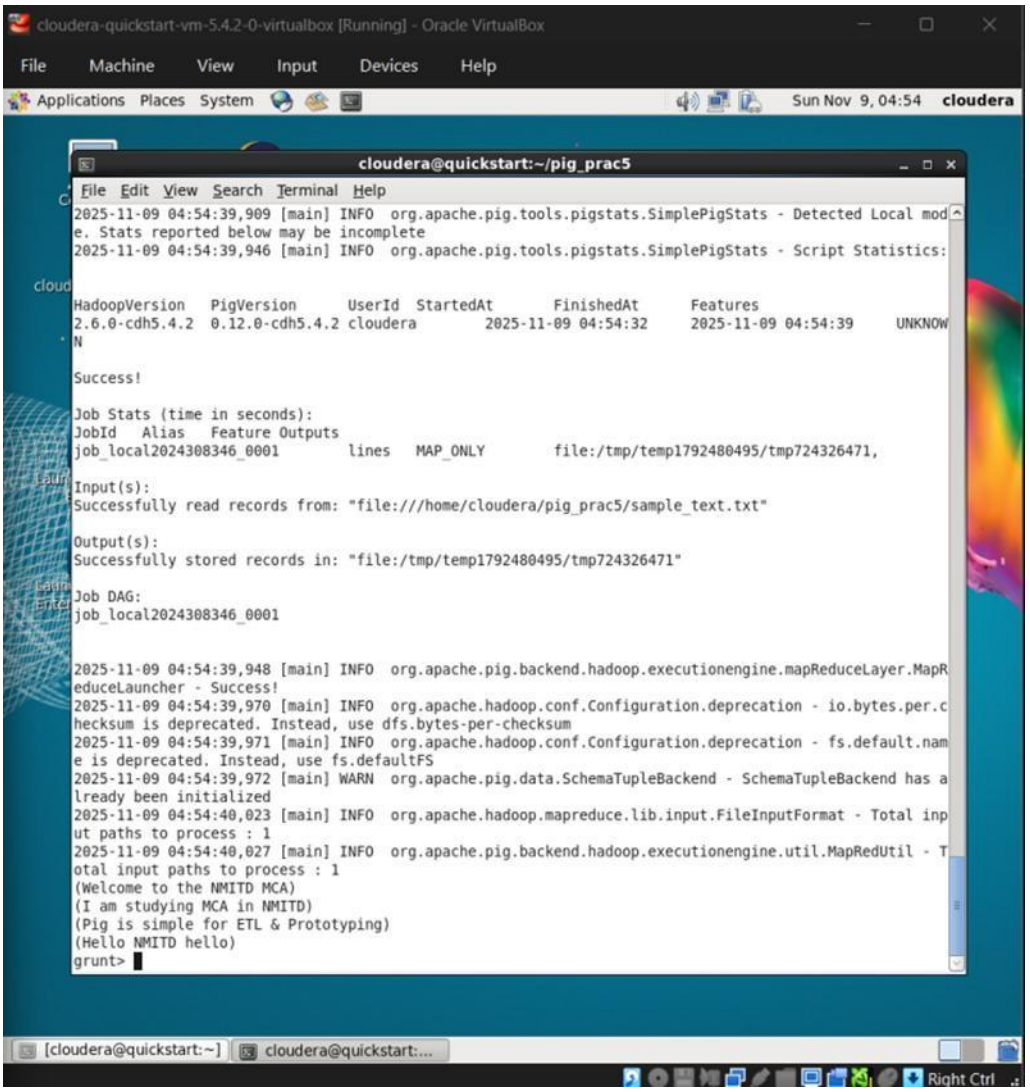

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Nov 9, 04:50 cloudera
cloudera@quickstart:~/pig_prac5
File Edit View Search Terminal Help
[cloudera@quickstart pig_prac5]$ pig -x local
log4j:WARN No appenders could be found for logger (org.apache.hadoop.util.Shell)
.
log4j:WARN Please initialize the log4j system properly.
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more in
fo.
2025-11-09 04:48:27,170 [main] INFO org.apache.pig.Main - Apache Pig version 0.
12.0-cdh5.4.2 (rexported) compiled May 19 2015, 17:03:41
2025-11-09 04:48:27,186 [main] INFO org.apache.pig.Main - Logging error message
s to: /home/cloudera/pig_prac5/pig_1762692506852.log
2025-11-09 04:48:27,567 [main] INFO org.apache.pig.impl.util.Utils - Default bo
otup file /home/cloudera/.pigbootup not found
2025-11-09 04:48:31,625 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2025-11-09 04:48:31,626 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - mapred.job.tracker is deprecated. Instead, use mapreduce.jobtracker.addr
ess
2025-11-09 04:48:31,633 [main] INFO org.apache.pig.backend.hadoop.executionengi
ne.HExecutionEngine - Connecting to hadoop file system at: file:///
2025-11-09 04:48:35,148 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-11-09 04:48:35,156 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2025-11-09 04:48:35,445 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2025-11-09 04:48:35,460 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-11-09 04:48:35,729 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2025-11-09 04:48:35,747 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-11-09 04:48:36,170 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2025-11-09 04:48:36,178 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-11-09 04:48:36,363 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
2025-11-09 04:48:36,371 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2025-11-09 04:48:36,508 [main] INFO org.apache.hadoop.conf.Configuration.deprec
ation - fs.default.name is deprecated. Instead, use fs.defaultFS
[cloudera@quickstart:~]
cloudera@quickstart:...

BIGDATA



Load and view data

```
ation - io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> lines = LOAD 'sample_text.txt' AS (line:chararray);
grunt> DUMP lines;
2025-11-09 04:54:29,083 [main] INFO  org.apache.pig.tools.pigstats.ScriptState - Pig features used in
```



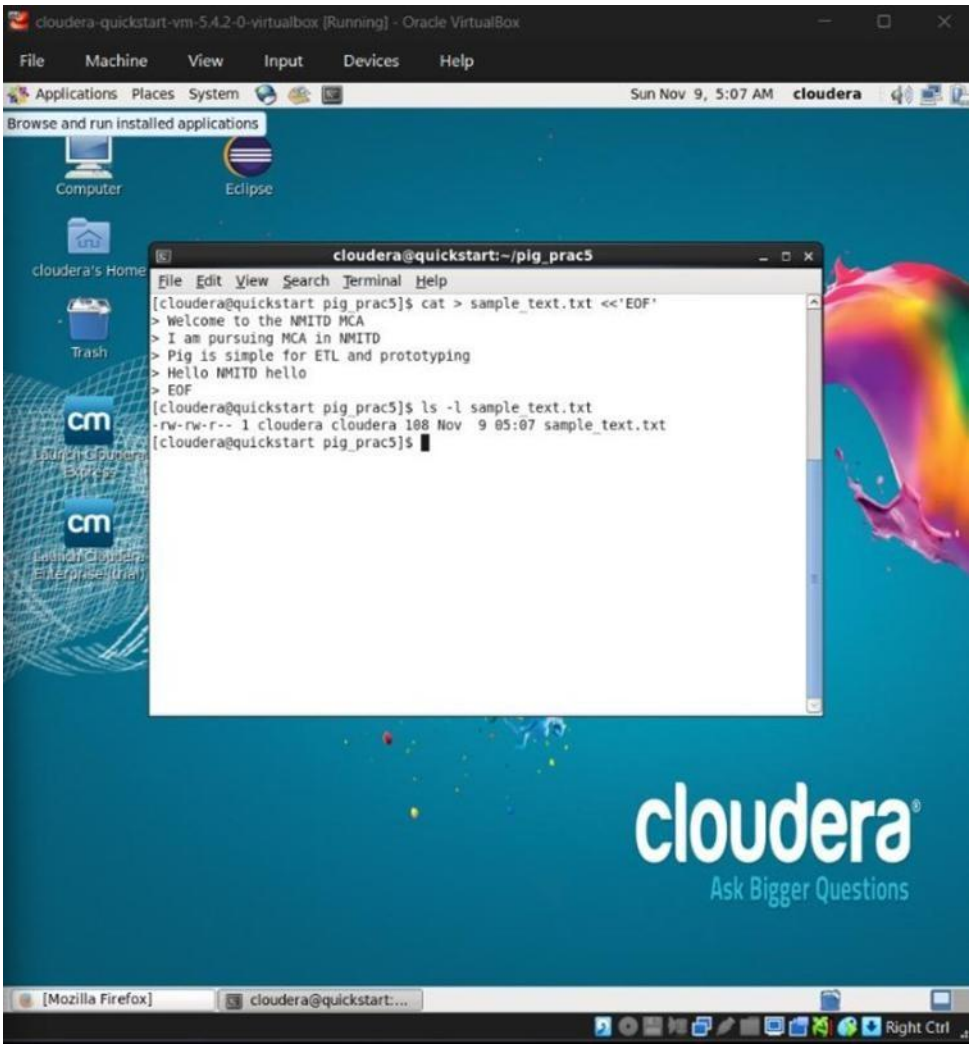
Write and Execute WordCount Script in Local Mode

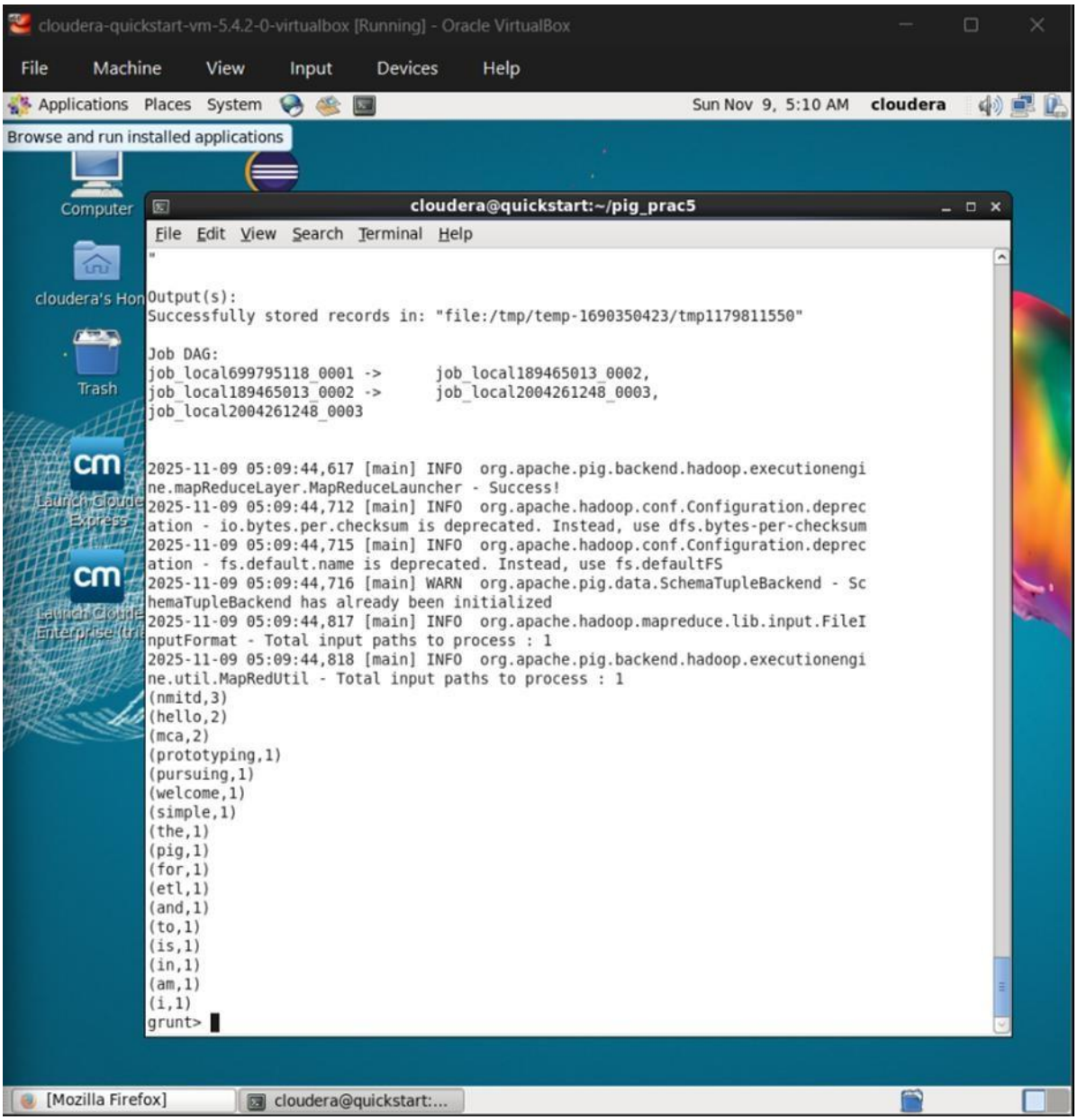
Exit Grunt shell using:

quite;

Create script file (Write Pig Latin script for word count in file):

BIGDATA

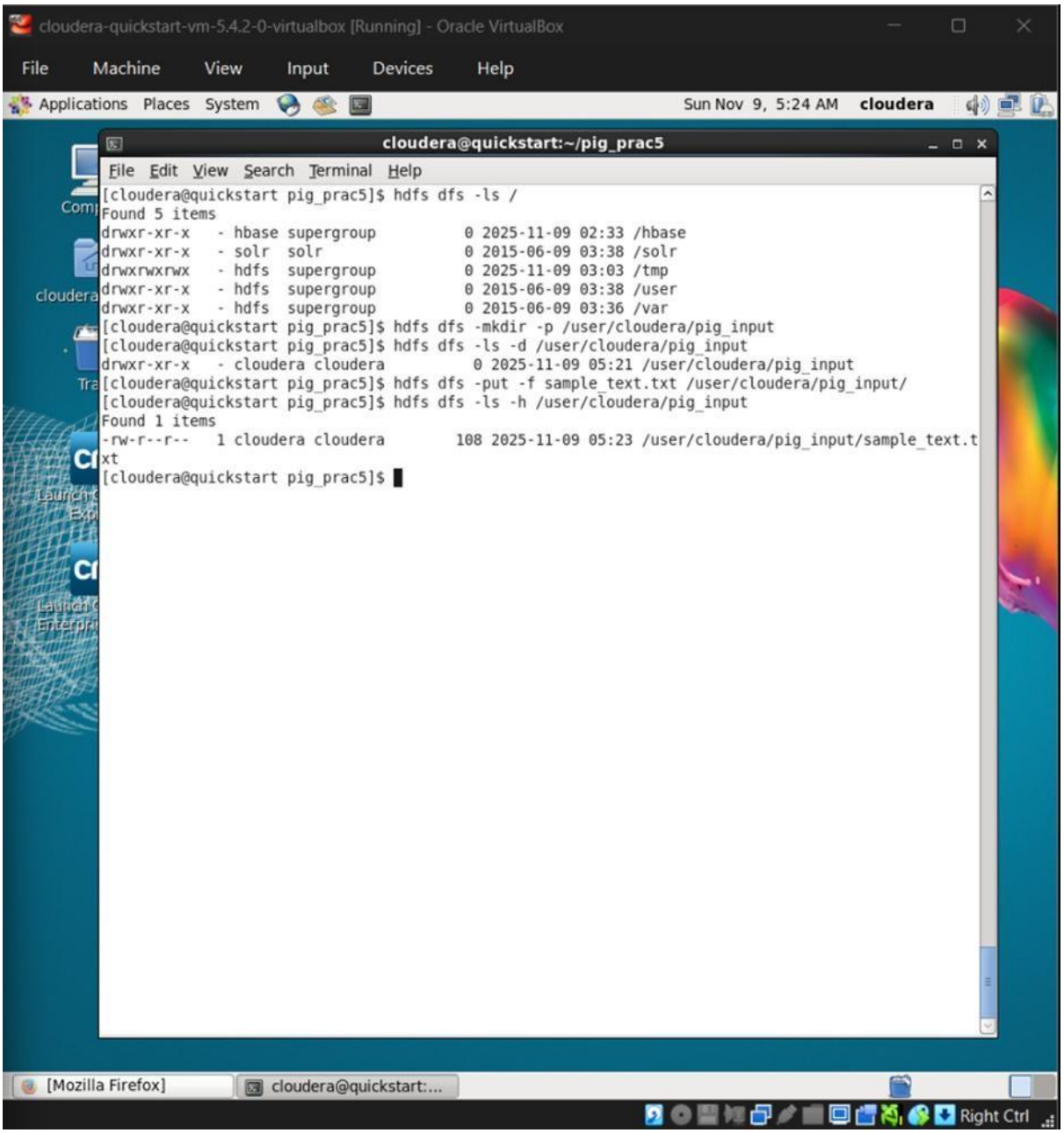

cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Nov 9, 5:07 AM cloudera
Browse and run installed applications
Computer
Eclipse
cloudera's Home
Trash
cloudera@quickstart:~/pig_prac5
File Edit View Search Terminal Help
[cloudera@quickstart pig_prac5]$ cat > sample_text.txt <<'EOF'
> Welcome to the NMITD MCA
> I am pursuing MCA in NMITD
> Pig is simple for ETL and prototyping
> Hello NMITD hello
> EOF
[cloudera@quickstart pig_prac5]$ ls -l sample_text.txt
-rw-rw-r-- 1 cloudera cloudera 108 Nov 9 05:07 sample_text.txt
[cloudera@quickstart pig_prac5]$
cloudera
Ask Bigger Questions
[Mozilla Firefox]
cloudera@quickstart:...
Right Ctrl

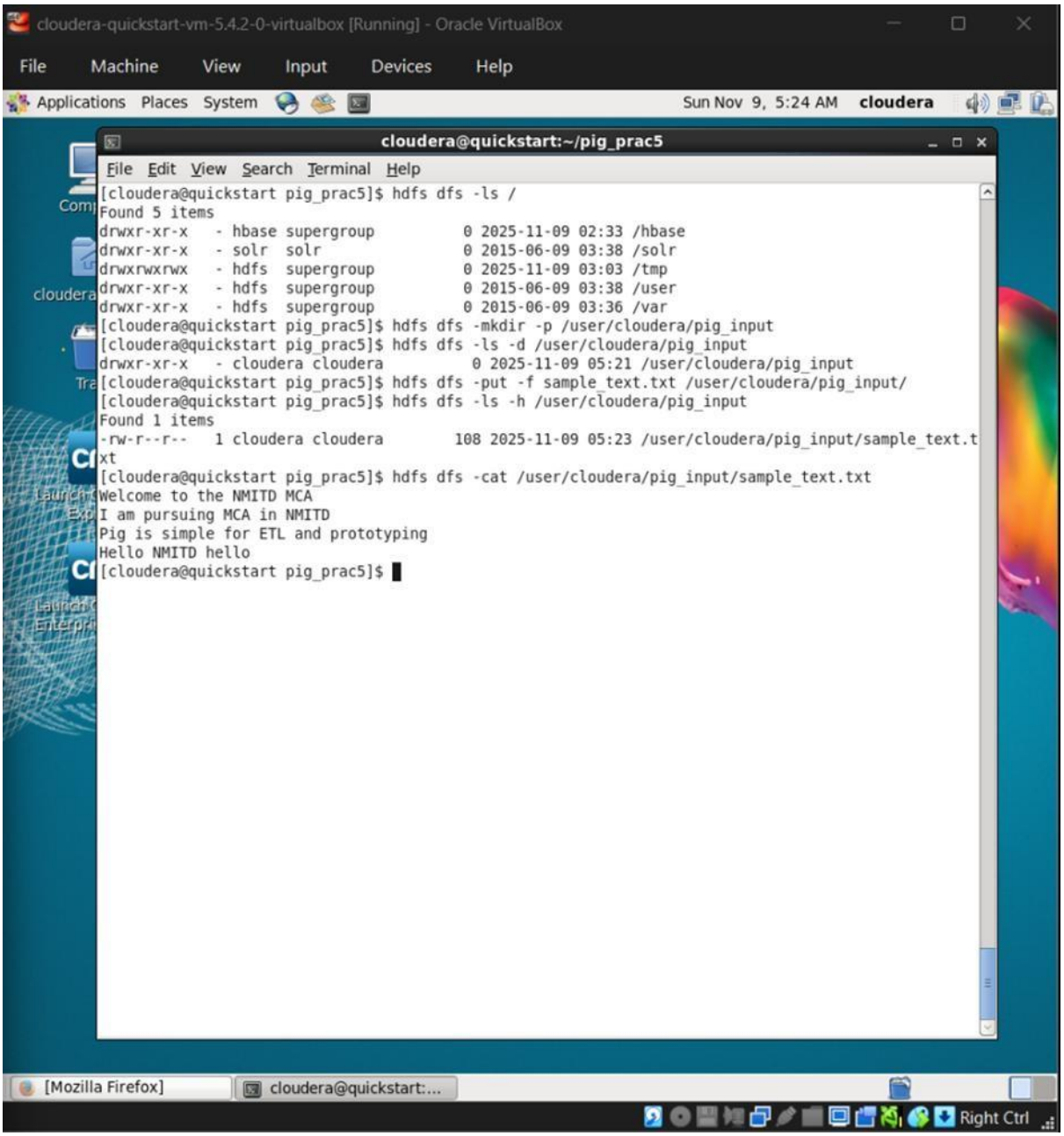


Prepare Data in HDFS (for MapReduce Mode)

Create directory and upload file to HDFS for MapReduce

BIGDATA


cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Nov 9, 5:24 AM cloudera
cloudera@quickstart:~/pig_prac5
File Edit View Search Terminal Help
[cloudera@quickstart pig_prac5]$ hdfs dfs -ls /
Found 5 items
drwxr-xr-x - hbase supergroup 0 2025-11-09 02:33 /hbase
drwxr-xr-x - solr solr 0 2015-06-09 03:38 /solr
drwxrwxrwx - hdfs supergroup 0 2025-11-09 03:03 /tmp
drwxr-xr-x - hdfs supergroup 0 2015-06-09 03:38 /user
drwxr-xr-x - hdfs supergroup 0 2015-06-09 03:36 /var
[cloudera@quickstart pig_prac5]$ hdfs dfs -mkdir -p /user/cloudera/pig_input
[cloudera@quickstart pig_prac5]$ hdfs dfs -ls -d /user/cloudera/pig_input
drwxr-xr-x - cloudera cloudera 0 2025-11-09 05:21 /user/cloudera/pig_input
[cloudera@quickstart pig_prac5]$ hdfs dfs -put -f sample_text.txt /user/cloudera/pig_input/
[cloudera@quickstart pig_prac5]$ hdfs dfs -ls -h /user/cloudera/pig_input
Found 1 items
-rw-r--r-- 1 cloudera cloudera 108 2025-11-09 05:23 /user/cloudera/pig_input/sample_text.t
xt
[cloudera@quickstart pig_prac5]$
[Mozilla Firefox]
cloudera@quickstart:...
Right Ctrl

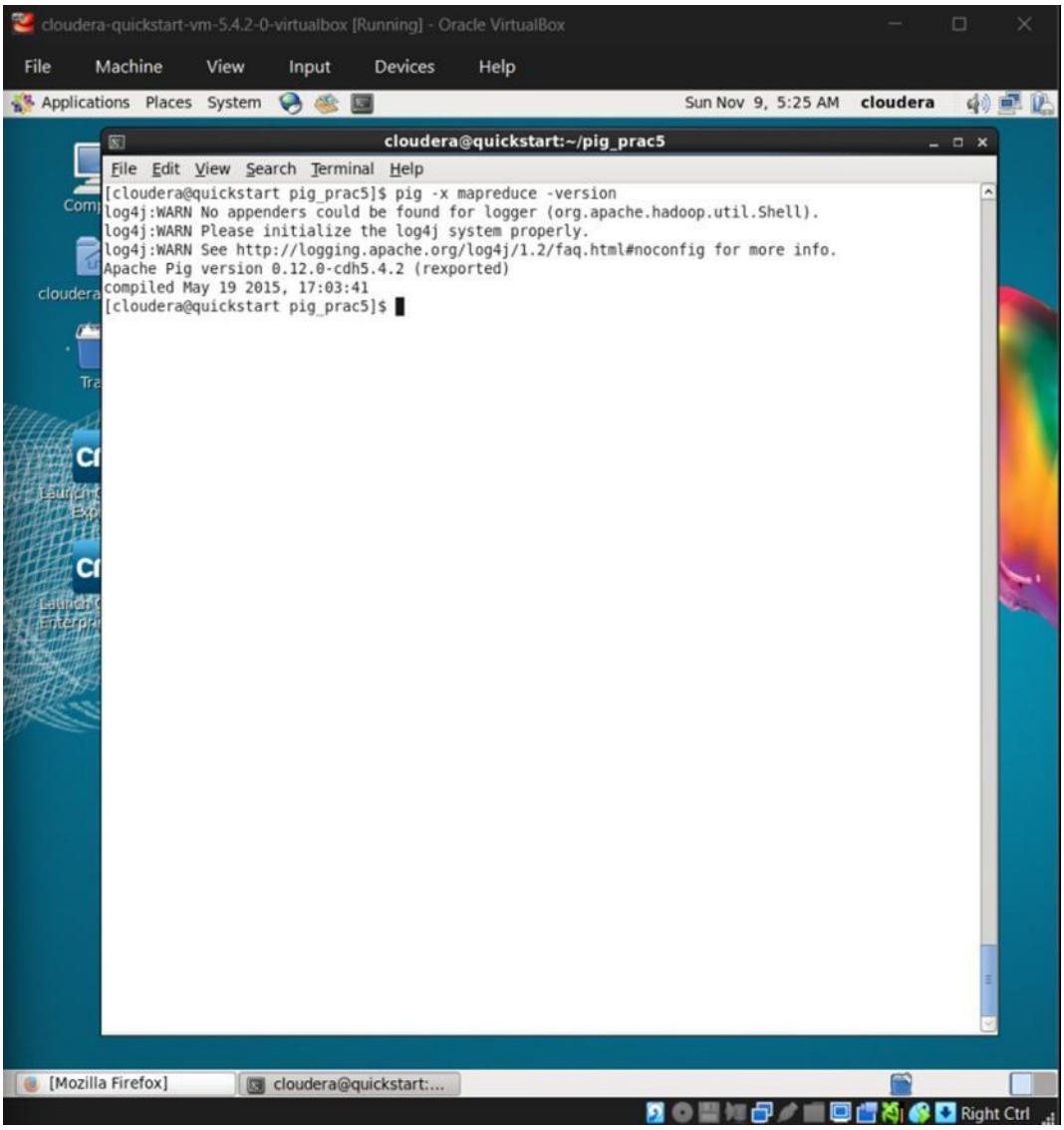


Verify Pig MapReduce Mode

Check Pig's MapReduce engine connectivity:

- pig -x mapreduce –version

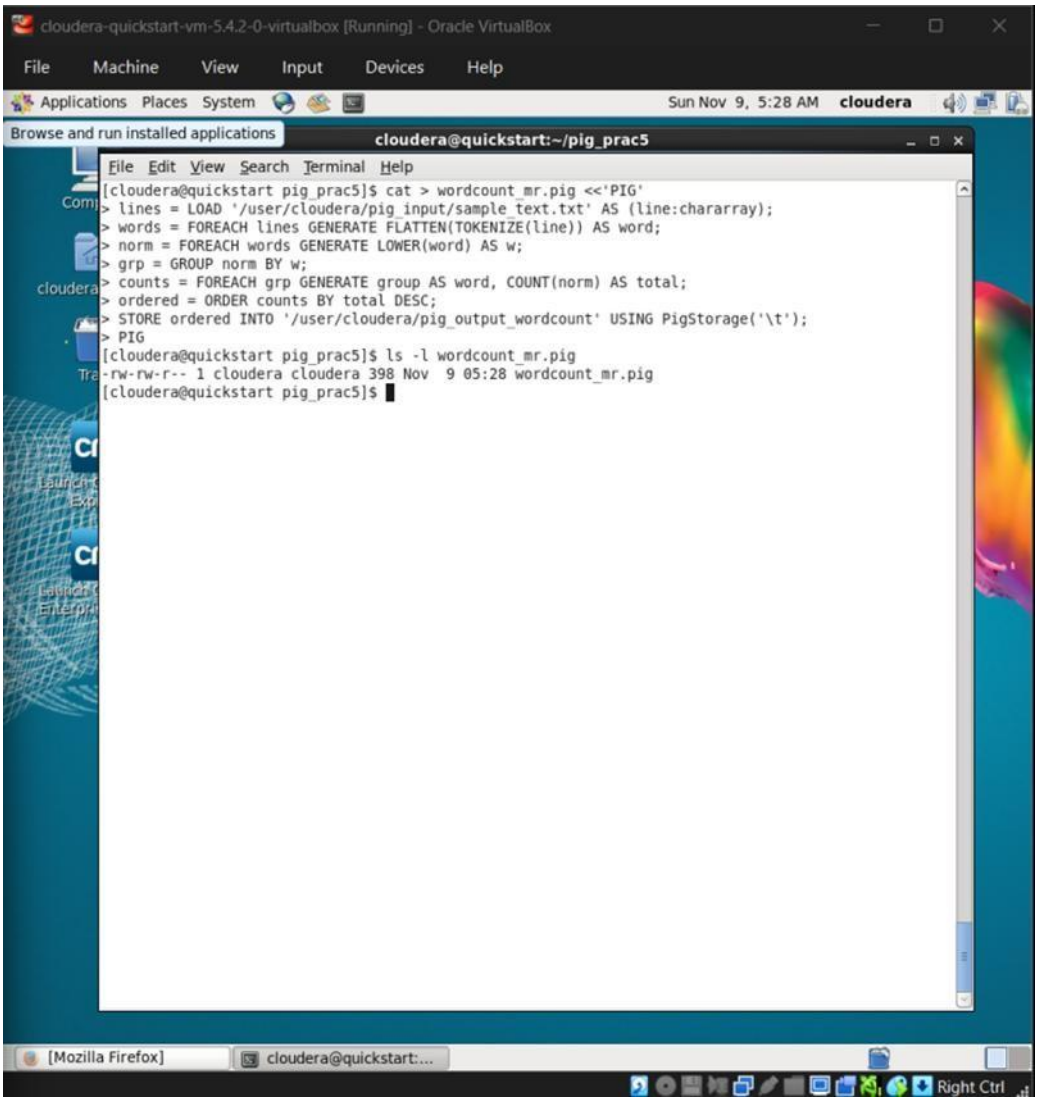
BIGDATA



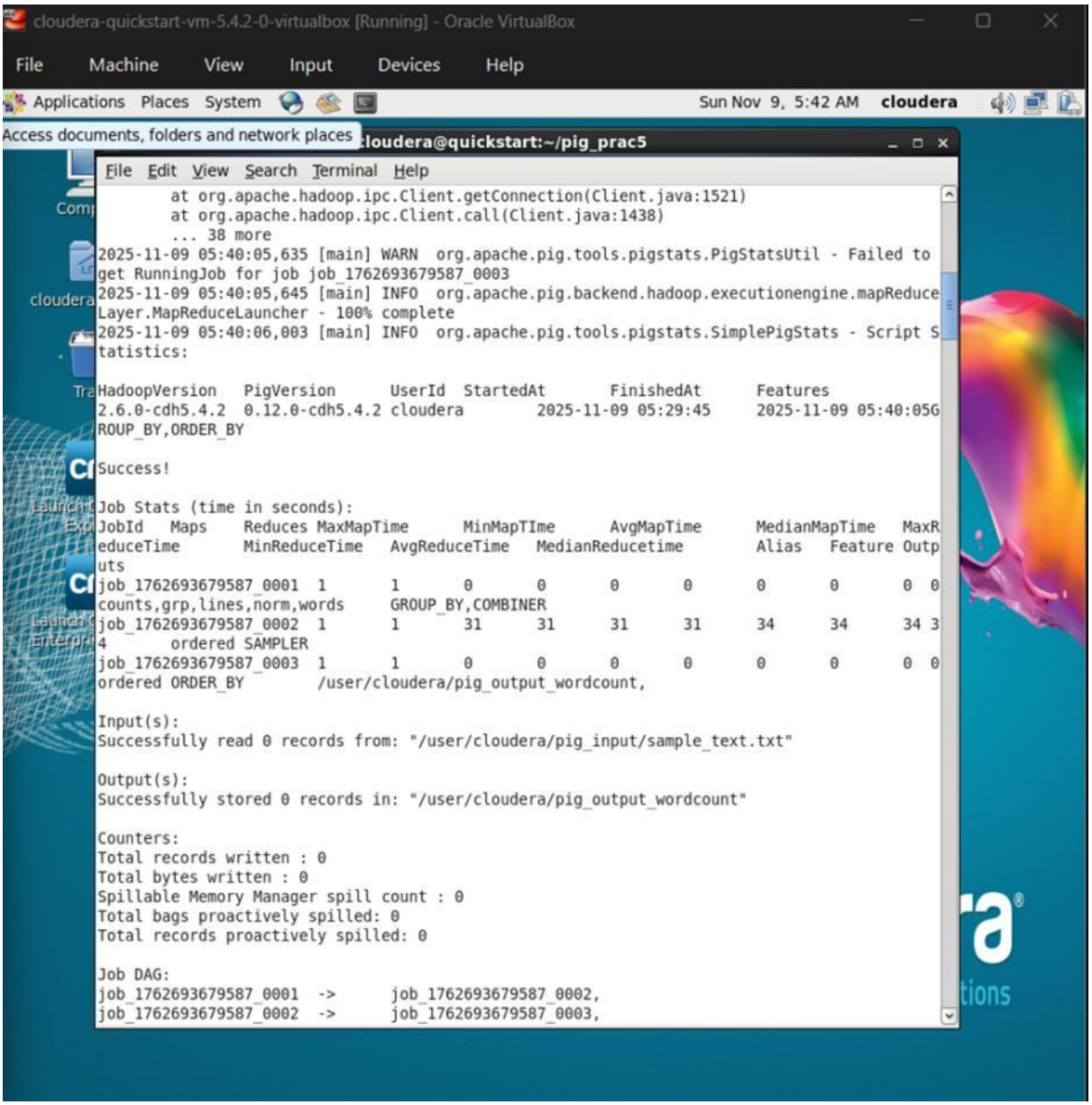
Create Pig WordCount Script for HDFS

Create the Pig script:

BIGDATA



Run Pig in MapReduce Mode
Execute the script
pig wordcount_mr.pig
Pig launches MapReduce jobs & output path displayed



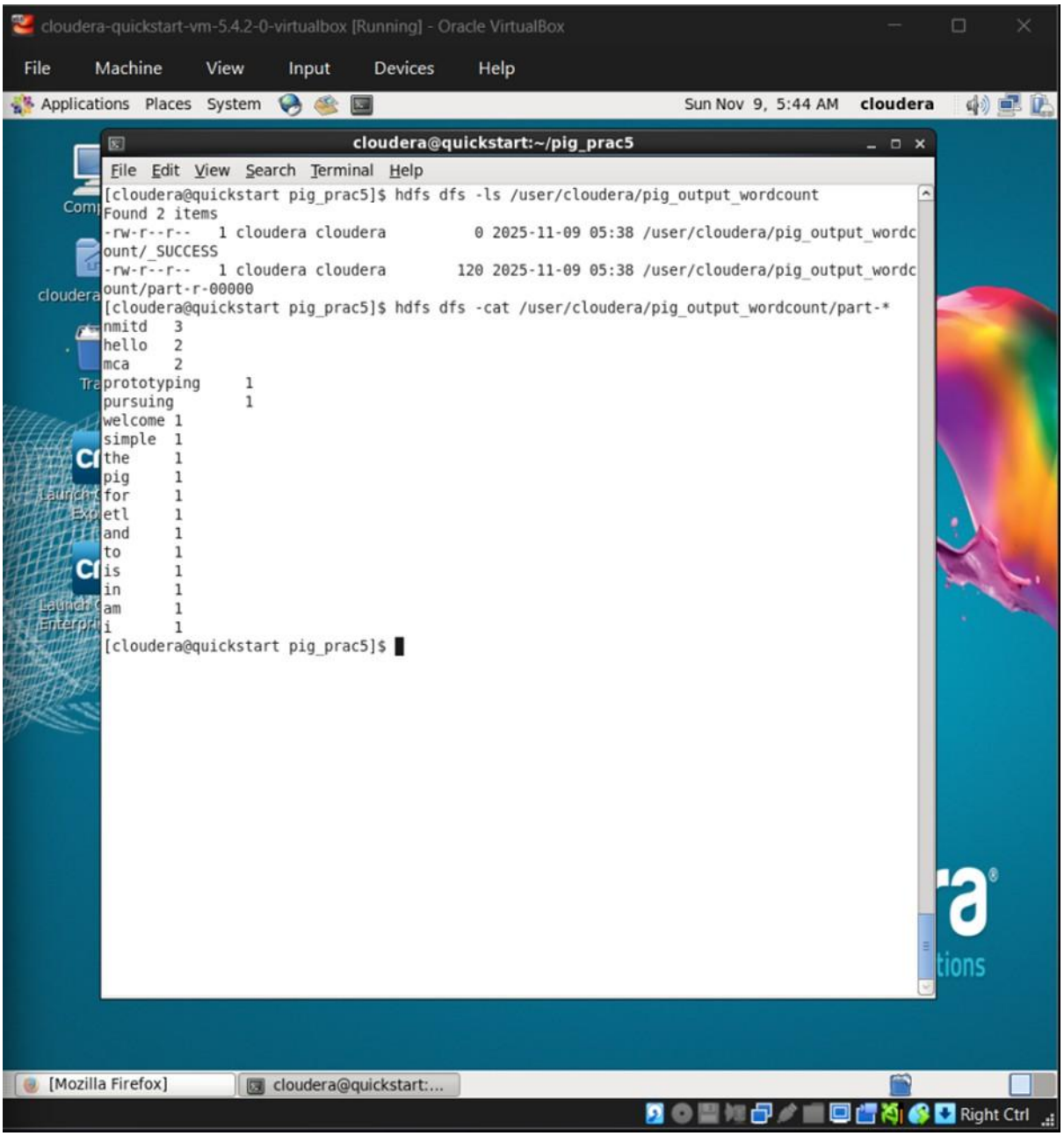
Verify the Output in HDFS

Check output directory

hdfs dfs -ls /user/cloudera/pig_output_wordcount

hdfs dfs -cat /user/cloudera/pig_output_wordcount/part-r-*

BIGDATA


cloudera-quickstart-vm-5.4.2-0-virtualbox [Running] - Oracle VirtualBox
File Machine View Input Devices Help
Applications Places System
Sun Nov 9, 5:44 AM cloudera
cloudera@quickstart:~/pig_prac5
File Edit View Search Terminal Help
[cloudera@quickstart pig_prac5]$ hdfs dfs -ls /user/cloudera/pig_output_wordcount
Found 2 items
-rw-r--r-- 1 cloudera cloudera 0 2025-11-09 05:38 /user/cloudera/pig_output_wordcount/_SUCCESS
-rw-r--r-- 1 cloudera cloudera 120 2025-11-09 05:38 /user/cloudera/pig_output_wordcount/part-r-00000
[cloudera@quickstart pig_prac5]$ hdfs dfs -cat /user/cloudera/pig_output_wordcount/part-*
nmitd 3
hello 2
mca 2
prototyping 1
pursuing 1
welcome 1
simple 1
the 1
pig 1
for 1
etl 1
and 1
to 1
is 1
in 1
am 1
i 1
[cloudera@quickstart pig_prac5]$
[Mozilla Firefox]
cloudera@quickstart:...
Right Ctrl

6. Spark:

Apache Spark Commands in Scala	Pair RDD (Key-Value RDD) Operations
Start Spark Shell	Create Pair RDD
Create RDD from Collection	reduceByKey
Read Text File into RDD (from HDFS)	groupByKey — With Output Viewing
Map Transformation	mapValues
Filter Transformation	sortByKey
Reduce Action	join
Collect Action	cogroup
Save RDD to HDFS	aggregateByKey
Cache/Persist RDD	foldByKey
	Read from Local File

Start Spark Shell:

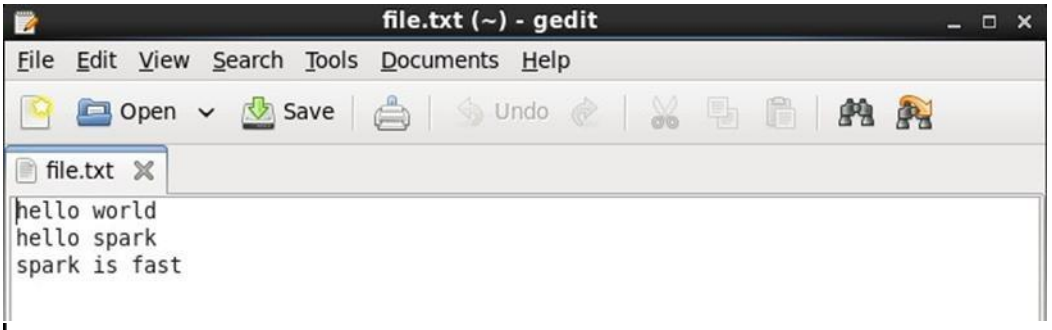
```
[cloudera@quickstart ~]$ spark-shell --master local[*]
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/jars/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/parquet/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
25/11/05 23:52:14 INFO SecurityManager: Changing view acls to: cloudera
25/11/05 23:52:14 INFO SecurityManager: Changing modify acls to: cloudera
25/11/05 23:52:14 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(cloudera); users with modify permissions: Set(cloudera)
25/11/05 23:52:14 INFO HttpServer: Starting HTTP Server
25/11/05 23:52:15 INFO Server: jetty-8.y.z-SNAPSHOT
25/11/05 23:52:15 INFO AbstractConnector: Started SocketConnector@0.0.0.0:34077
25/11/05 23:52:15 INFO Utils: Successfully started service 'HTTP class server' on port 34077.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 1.3.0
      /_/
```

Create RDD from Collection:

```
scala> val numbers = Seq(1, 2, 3, 4, 5)
numbers: Seq[Int] = List(1, 2, 3, 4, 5)

scala> val rdd = sc.parallelize(numbers)
rdd: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[0] at parallelize at <console>:23
```

```
scala> rdd.collect()
25/11/06 00:46:13 INFO SparkContext: Starting job: collect at <console>:26
25/11/06 00:46:13 INFO DAGScheduler: Got job 0 (collect at <console>:26) with 2 output partitions (allowLocal=false)
25/11/06 00:46:13 INFO DAGScheduler: Final stage: Stage 0(collect at <console>:26)
25/11/06 00:46:13 INFO DAGScheduler: Parents of final stage: List()
25/11/06 00:46:13 INFO DAGScheduler: Missing parents: List()
25/11/06 00:46:13 INFO DAGScheduler: Submitting Stage 0 (ParallelCollectionRDD[0] at parallelize at <console>:23), which has no missing parents
25/11/06 00:46:14 INFO MemoryStore: ensureFreeSpace(1128) called with curMem=0, maxMem=278302556
25/11/06 00:46:14 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 1128.0 B, free 265.4 MB)
25/11/06 00:46:15 INFO MemoryStore: ensureFreeSpace(800) called with curMem=1128, maxMem=278302556
25/11/06 00:46:15 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 800.0 B, free 265.4 MB)
25/11/06 00:46:15 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on quickstart.cloudera:54520 (size: 800.0 B, free: 265.4 MB)
25/11/06 00:46:15 INFO BlockManagerMaster: Updated info of block broadcast_0_piece0
25/11/06 00:46:15 INFO SparkContext: Created broadcast 0 from broadcast at DAGScheduler.scala:839
25/11/06 00:46:15 INFO DAGScheduler: Submitting 2 missing tasks from Stage 0 (ParallelCollectionRDD[0] at parallelize at <console>:23)
25/11/06 00:46:15 INFO YarnScheduler: Adding task set 0.0 with 2 tasks
25/11/06 00:46:15 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, quickstart.cloudera, PROCESS_LOCAL, 1162 bytes)
25/11/06 00:46:18 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on quickstart.cloudera:44681 (size: 800.0 B, free: 530.3 MB)
25/11/06 00:46:19 INFO TaskSetManager: Starting task 1.0 in stage 0.0 (TID 1, quickstart.cloudera, PROCESS_LOCAL, 1163 bytes)
25/11/06 00:46:19 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 3765 ms on quickstart.cloudera (1/2)
25/11/06 00:46:19 INFO TaskSetManager: Finished task 1.0 in stage 0.0 (TID 1) in 140 ms on quickstart.cloudera (2/2)
25/11/06 00:46:19 INFO DAGScheduler: Stage 0 (collect at <console>:26) finished in 3.862 s
25/11/06 00:46:19 INFO YarnScheduler: Removed TaskSet 0.0, whose tasks have all completed, from pool
25/11/06 00:46:19 INFO DAGScheduler: Job 0 finished: collect at <console>:26, took 5.641321 s
res0: Array[Int] = Array(1, 2, 3, 4, 5)
```

Read Text File into RDD:




```
scala> val textRdd = sc.textFile("file:/home/cloudera/file.txt")
25/11/06 00:46:48 INFO BlockManager: Removing broadcast 0
25/11/06 00:46:48 INFO BlockManager: Removing block broadcast_0
25/11/06 00:46:48 INFO MemoryStore: Block broadcast_0 of size 1128 dropped from memory (free 278301756)
25/11/06 00:46:48 INFO BlockManager: Removing block broadcast_0_piece0
25/11/06 00:46:48 INFO MemoryStore: Block broadcast_0_piece0 of size 800 dropped from memory (free 278302556)
25/11/06 00:46:48 INFO BlockManagerInfo: Removed broadcast_0_piece0 on quickstart.cloudera:54520 in memory (size: 800.0 B, free: 265.4 MB)
25/11/06 00:46:48 INFO BlockManagerMaster: Updated info of block broadcast_0_piece0
25/11/06 00:46:48 INFO BlockManagerInfo: Removed broadcast_0_piece0 on quickstart.cloudera:44681 in memory (size: 800.0 B, free: 530.3 MB)
25/11/06 00:46:48 INFO ContextCleaner: Cleaned broadcast 0
25/11/06 00:46:48 INFO MemoryStore: ensureFreeSpace(283020) called with curMem=0, maxMem=278302556
25/11/06 00:46:48 INFO MemoryStore: Block broadcast_1 stored as values in memory (estimated size 276.4 KB, free 265.1 MB)
25/11/06 00:46:48 INFO MemoryStore: ensureFreeSpace(22296) called with curMem=283020, maxMem=278302556
25/11/06 00:46:48 INFO MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 21.8 KB, free 265.1 MB)
25/11/06 00:46:48 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on quickstart.cloudera:54520 (size: 21.8 KB, free: 265.4 MB)
25/11/06 00:46:48 INFO BlockManagerMaster: Updated info of block broadcast_1_piece0
25/11/06 00:46:48 INFO SparkContext: Created broadcast 1 from textFile at <console>:21
textRdd: org.apache.spark.rdd.RDD[String] = file:/home/cloudera/file.txt MapPartitionsRDD[2] at textFile at <console>:21

scala> textRdd.collect().foreach(println)
25/11/06 00:47:01 INFO FileInputFormat: Total input paths to process : 1
25/11/06 00:47:01 INFO SparkContext: Starting job: collect at <console>:24
25/11/06 00:47:01 INFO DAGScheduler: Got job 1 (collect at <console>:24) with 2 output partitions (allowLocal=false)
25/11/06 00:47:01 INFO DAGScheduler: Final stage: Stage 1(collect at <console>:24)
25/11/06 00:47:01 INFO DAGScheduler: Parents of final stage: List()
25/11/06 00:47:01 INFO DAGScheduler: Missing parents: List()
25/11/06 00:47:01 INFO DAGScheduler: Submitting Stage 1 (file:/home/cloudera/file.txt MapPartitionsRDD[2] at textFile at <console>:21), which has no missing parents
25/11/06 00:47:01 INFO MemoryStore: ensureFreeSpace(2664) called with curMem=305316, maxMem=278302556
25/11/06 00:47:01 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 2.6 KB, free 265.1 MB)
25/11/06 00:47:01 INFO MemoryStore: ensureFreeSpace(1634) called with curMem=307980, maxMem=278302556
25/11/06 00:47:01 INFO MemoryStore: Block broadcast_2_piece0 stored as bytes in memory (estimated size 1634.0 B, free 265.1 MB)
25/11/06 00:47:01 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on quickstart.cloudera:54520 (size: 1634.0 B, free: 265.4 MB)
25/11/06 00:47:01 INFO BlockManagerMaster: Updated info of block broadcast_2_piece0
25/11/06 00:47:01 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:839
25/11/06 00:47:01 INFO DAGScheduler: Submitting 2 missing tasks from Stage 1 (file:/home/cloudera/file.txt MapPartitionsRDD[2] at textFile at <console>:21)
25/11/06 00:47:01 INFO YarnScheduler: Adding task set 1.0 with 2 tasks
25/11/06 00:47:01 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 2, quickstart.cloudera, PROCESS_LOCAL, 1292 bytes)
25/11/06 00:47:01 INFO BlockManagerInfo: Added broadcast_2_piece0 in memory on quickstart.cloudera:44681 (size: 1634.0 B, free: 530.3 MB)
25/11/06 00:47:01 INFO BlockManagerInfo: Added broadcast_1_piece0 in memory on quickstart.cloudera:44681 (size: 21.8 KB, free: 530.3 MB)
25/11/06 00:47:06 INFO TaskSetManager: Starting task 1.0 in stage 1.0 (TID 3, quickstart.cloudera, PROCESS_LOCAL, 1292 bytes)
25/11/06 00:47:06 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 2) in 5275 ms on quickstart.cloudera (1/2)
25/11/06 00:47:06 INFO DAGScheduler: Stage 1 (collect at <console>:24) finished in 5.335 s
25/11/06 00:47:06 INFO DAGScheduler: Job 1 finished: collect at <console>:24, took 5.386064 s
hello world
hello spark
spark is fast
```

Map Transformation:

```
scala> val squares = rdd.map(x => x * x)
25/11/06 00:51:05 INFO BlockManager: Removing broadcast 2
25/11/06 00:51:05 INFO BlockManager: Removing block broadcast_2
25/11/06 00:51:05 INFO MemoryStore: Block broadcast_2 of size 2664 dropped from memory (free 277995606)
25/11/06 00:51:05 INFO BlockManager: Removing block broadcast_2_piece0
25/11/06 00:51:05 INFO MemoryStore: Block broadcast_2_piece0 of size 1634 dropped from memory (free 277997240)
25/11/06 00:51:05 INFO BlockManagerInfo: Removed broadcast_2_piece0 on quickstart.cloudera:54520 in memory (size: 1634.0 B, free: 265.4 MB)
25/11/06 00:51:05 INFO BlockManagerMaster: Updated info of block broadcast_2_piece0
25/11/06 00:51:05 INFO BlockManagerInfo: Removed broadcast_2_piece0 on quickstart.cloudera:44681 in memory (size: 1634.0 B, free: 530.3 MB)
25/11/06 00:51:05 INFO ContextCleaner: Cleaned broadcast 2
squares: org.apache.spark.rdd.RDD[Int] = MapPartitionsRDD[3] at map at <console>:25

scala> squares.collect()
25/11/06 00:51:17 INFO SparkContext: Starting job: collect at <console>:28
25/11/06 00:51:17 INFO DAGScheduler: Got job 2 (collect at <console>:28) with 2 output partitions (allowLocal=false)
25/11/06 00:51:17 INFO DAGScheduler: Final stage: Stage 2(collect at <console>:28)
25/11/06 00:51:17 INFO DAGScheduler: Parents of final stage: List()
25/11/06 00:51:17 INFO DAGScheduler: Missing parents: List()
25/11/06 00:51:17 INFO DAGScheduler: Submitting Stage 2 (MapPartitionsRDD[3] at map at <console>:25), which has no missing parents
25/11/06 00:51:17 INFO MemoryStore: ensureFreeSpace(1792) called with curMem=305316, maxMem=278302556
25/11/06 00:51:17 INFO MemoryStore: Block broadcast_3 stored as values in memory (estimated size 1792.0 B, free 265.1 MB)
25/11/06 00:51:17 INFO MemoryStore: ensureFreeSpace(1157) called with curMem=307108, maxMem=278302556
25/11/06 00:51:17 INFO MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 1157.0 B, free 265.1 MB)
25/11/06 00:51:17 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on quickstart.cloudera:54520 (size: 1157.0 B, free: 265.4 MB)
25/11/06 00:51:17 INFO BlockManagerMaster: Updated info of block broadcast_3_piece0
25/11/06 00:51:17 INFO SparkContext: Created broadcast 3 from broadcast at DAGScheduler.scala:839
25/11/06 00:51:17 INFO DAGScheduler: Submitting 2 missing tasks from Stage 2 (MapPartitionsRDD[3] at map at <console>:25)
25/11/06 00:51:17 INFO YarnScheduler: Adding task set 2.0 with 2 tasks
25/11/06 00:51:17 INFO TaskSetManager: Starting task 0.0 in stage 2.0 (TID 4, quickstart.cloudera, PROCESS_LOCAL, 1162 bytes)
25/11/06 00:51:17 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on quickstart.cloudera:44681 (size: 1157.0 B, free: 530.3 MB)
25/11/06 00:51:17 INFO TaskSetManager: Starting task 1.0 in stage 2.0 (TID 5, quickstart.cloudera, PROCESS_LOCAL, 1163 bytes)
25/11/06 00:51:17 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 4) in 422 ms on quickstart.cloudera (1/2)
res2: Array[Int] = Array(1, 4, 9, 16, 25)
```

Filter Transformation:

```
scala> val even = rdd.filter(x => x % 2 == 0)
even: org.apache.spark.rdd.RDD[Int] = MapPartitionsRDD[4] at filter at <console>:25

scala> even.collect()
25/11/06 00:51:37 INFO SparkContext: Starting job: collect at <console>:28
25/11/06 00:51:37 INFO DAGScheduler: Got job 3 (collect at <console>:28) with 2 output partitions (allowLocal=false)
25/11/06 00:51:37 INFO DAGScheduler: Final stage: Stage 3(collect at <console>:28)
25/11/06 00:51:37 INFO DAGScheduler: Parents of final stage: List()
25/11/06 00:51:37 INFO DAGScheduler: Missing parents: List()
25/11/06 00:51:37 INFO DAGScheduler: Submitting Stage 3 (MapPartitionsRDD[4] at filter at <console>:25), which has no missing parents
25/11/06 00:51:37 INFO MemoryStore: ensureFreeSpace(1792) called with curMem=308265, maxMem=278302556
25/11/06 00:51:37 INFO MemoryStore: Block broadcast_4 stored as values in memory (estimated size 1792.0 B, free 265.1 MB)
25/11/06 00:51:37 INFO MemoryStore: ensureFreeSpace(1156) called with curMem=310057, maxMem=278302556
25/11/06 00:51:37 INFO MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (estimated size 1156.0 B, free 265.1 MB)
25/11/06 00:51:37 INFO BlockManagerInfo: Added broadcast_4_piece0 in memory on quickstart.cloudera:54520 (size: 1156.0 B, free: 265.4 MB)
25/11/06 00:51:37 INFO BlockManagerMaster: Updated info of block broadcast_4_piece0
25/11/06 00:51:37 INFO SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:839
25/11/06 00:51:37 INFO DAGScheduler: Submitting 2 missing tasks from Stage 3 (MapPartitionsRDD[4] at filter at <console>:25)
25/11/06 00:51:37 INFO YarnScheduler: Adding task set 3.0 with 2 tasks
25/11/06 00:51:37 INFO TaskSetManager: Starting task 0.0 in stage 3.0 (TID 6, quickstart.cloudera, PROCESS_LOCAL, 1162 bytes)
25/11/06 00:51:37 INFO BlockManagerInfo: Added broadcast_4_piece0 in memory on quickstart.cloudera:44681 (size: 1156.0 B, free: 530.3 MB)
25/11/06 00:51:37 INFO TaskSetManager: Starting task 1.0 in stage 3.0 (TID 7, quickstart.cloudera, PROCESS_LOCAL, 1163 bytes)
25/11/06 00:51:37 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 6) in 156 ms on quickstart.cloudera (1/2)
25/11/06 00:51:37 INFO DAGScheduler: Stage 3 (collect at <console>:28) finished in 0.210 s
25/11/06 00:51:37 INFO DAGScheduler: Job 3 finished: collect at <console>:28, took 0.244475 s
res3: Array[Int] = Array(2, 4)
```

Reduce Action:

```
scala> val total = rdd.reduce((a, b) => a + b)
25/11/06 00:53:05 INFO SparkContext: Starting job: reduce at <console>:25
25/11/06 00:53:05 INFO DAGScheduler: Got job 4 (reduce at <console>:25) with 2 output partitions (allowLocal=false)
25/11/06 00:53:05 INFO DAGScheduler: Final stage: Stage 4(reduce at <console>:25)
25/11/06 00:53:05 INFO DAGScheduler: Parents of final stage: List()
25/11/06 00:53:05 INFO DAGScheduler: Missing parents: List()
25/11/06 00:53:05 INFO DAGScheduler: Submitting Stage 4 (ParallelCollectionRDD[0] at parallelize at <console>:23), which has no missing parents
25/11/06 00:53:05 INFO MemoryStore: ensureFreeSpace(1200) called with curMem=308264, maxMem=278302556
25/11/06 00:53:05 INFO MemoryStore: Block broadcast_5 stored as values in memory (estimated size 1200.0 B, free 265.1 MB)
25/11/06 00:53:05 INFO MemoryStore: ensureFreeSpace(832) called with curMem=309464, maxMem=278302556
25/11/06 00:53:05 INFO MemoryStore: Block broadcast_5_piece0 stored as bytes in memory (estimated size 832.0 B, free 265.1 MB)
25/11/06 00:53:05 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on quickstart.cloudera:54520 (size: 832.0 B, free: 265.4 MB)
25/11/06 00:53:05 INFO BlockManagerMaster: Updated info of block broadcast_5_piece0
25/11/06 00:53:05 INFO SparkContext: Created broadcast 5 from broadcast at DAGScheduler.scala:839
25/11/06 00:53:05 INFO DAGScheduler: Submitting 2 missing tasks from Stage 4 (ParallelCollectionRDD[0] at parallelize at <console>:23)
25/11/06 00:53:05 INFO YarnScheduler: Adding task set 4.0 with 2 tasks
25/11/06 00:53:05 INFO TaskSetManager: Starting task 0.0 in stage 4.0 (TID 8, quickstart.cloudera, PROCESS_LOCAL, 1162 bytes)
25/11/06 00:53:05 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on quickstart.cloudera:44681 (size: 832.0 B, free: 530.3 MB)
25/11/06 00:53:05 INFO TaskSetManager: Starting task 1.0 in stage 4.0 (TID 9, quickstart.cloudera, PROCESS_LOCAL, 1163 bytes)
25/11/06 00:53:05 INFO TaskSetManager: Finished task 0.0 in stage 4.0 (TID 8) in 331 ms on quickstart.cloudera (1/2)
25/11/06 00:53:05 INFO DAGScheduler: Stage 4 (reduce at <console>:25) finished in 0.455 s
25/11/06 00:53:05 INFO DAGScheduler: Job 4 finished: reduce at <console>:25, took 0.524725 s
total: Int = 15

scala> 25/11/06 00:53:05 INFO TaskSetManager: Finished task 1.0 in stage 4.0 (TID 9) in 122 ms on quickstart.cloudera (2/2)
25/11/06 00:53:05 INFO YarnScheduler: Removed TaskSet 4.0, whose tasks have all completed, from pool


scala> println(total)
15
```

Collect Action:

```
scala> rdd.collect().foreach(println)
25/11/06 00:53:31 INFO BlockManager: Removing broadcast 5
25/11/06 00:53:31 INFO BlockManager: Removing block broadcast_5_piece0
25/11/06 00:53:31 INFO MemoryStore: Block broadcast_5_piece0 of size 832 dropped from memory (free 277993092)
25/11/06 00:53:31 INFO BlockManagerInfo: Removed broadcast_5_piece0 on quickstart.cloudera:54520 in memory (size: 832.0 B, free: 265.4 MB)
25/11/06 00:53:31 INFO BlockManagerInfo: Removed broadcast_5_piece0 on quickstart.cloudera:44681 in memory (size: 832.0 B, free: 530.3 MB)
25/11/06 00:53:31 INFO BlockManagerMaster: Updated info of block broadcast_5_piece0
25/11/06 00:53:31 INFO BlockManager: Removing block broadcast_5
25/11/06 00:53:31 INFO MemoryStore: Block broadcast_5 of size 1200 dropped from memory (free 277994292)
25/11/06 00:53:31 INFO ContextCleaner: Cleaned broadcast 5
25/11/06 00:53:31 INFO SparkContext: Starting job: collect at <console>:26
25/11/06 00:53:31 INFO DAGScheduler: Got job 5 (collect at <console>:26) with 2 output partitions (allowLocal=false)
25/11/06 00:53:31 INFO DAGScheduler: Final stage: Stage 5(collect at <console>:26)
25/11/06 00:53:31 INFO DAGScheduler: Parents of final stage: List()
25/11/06 00:53:31 INFO DAGScheduler: Missing parents: List()
25/11/06 00:53:31 INFO DAGScheduler: Submitting Stage 5 (ParallelCollectionRDD[0] at parallelize at <console>:23), which has no missing parents
25/11/06 00:53:31 INFO MemoryStore: ensureFreeSpace(1128) called with curMem=308264, maxMem=278302556
25/11/06 00:53:31 INFO MemoryStore: Block broadcast_6 stored as values in memory (estimated size 1128.0 B, free 265.1 MB)
25/11/06 00:53:31 INFO MemoryStore: ensureFreeSpace(800) called with curMem=309392, maxMem=278302556
25/11/06 00:53:31 INFO MemoryStore: Block broadcast_6_piece0 stored as bytes in memory (estimated size 800.0 B, free 265.1 MB)
25/11/06 00:53:31 INFO BlockManagerInfo: Added broadcast_6_piece0 in memory on quickstart.cloudera:54520 (size: 800.0 B, free: 265.4 MB)
25/11/06 00:53:31 INFO BlockManagerMaster: Updated info of block broadcast_6_piece0
25/11/06 00:53:31 INFO SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:839
25/11/06 00:53:31 INFO DAGScheduler: Submitting 2 missing tasks from Stage 5 (ParallelCollectionRDD[0] at parallelize at <console>:23)
25/11/06 00:53:31 INFO YarnScheduler: Adding task set 5.0 with 2 tasks
25/11/06 00:53:31 INFO TaskSetManager: Starting task 0.0 in stage 5.0 (TID 10, quickstart.cloudera, PROCESS_LOCAL, 1162 bytes)
25/11/06 00:53:31 INFO BlockManagerInfo: Added broadcast_6_piece0 in memory on quickstart.cloudera:44681 (size: 800.0 B, free: 530.3 MB)
25/11/06 00:53:31 INFO TaskSetManager: Starting task 1.0 in stage 5.0 (TID 11, quickstart.cloudera, PROCESS_LOCAL, 1163 bytes)
25/11/06 00:53:31 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (TID 10) in 106 ms on quickstart.cloudera (1/2)
25/11/06 00:53:31 INFO DAGScheduler: Stage 5 (collect at <console>:26) finished in 0.237 s
25/11/06 00:53:31 INFO DAGScheduler: Job 5 finished: collect at <console>:26, took 0.271782 s
25/11/06 00:53:31 INFO TaskSetManager: Finished task 1.0 in stage 5.0 (TID 11) in 135 ms on quickstart.cloudera (2/2)
25/11/06 00:53:31 INFO YarnScheduler: Removed TaskSet 5.0, whose tasks have all completed, from pool
1
2
3
4
5
```

Save RDD to HDFS:

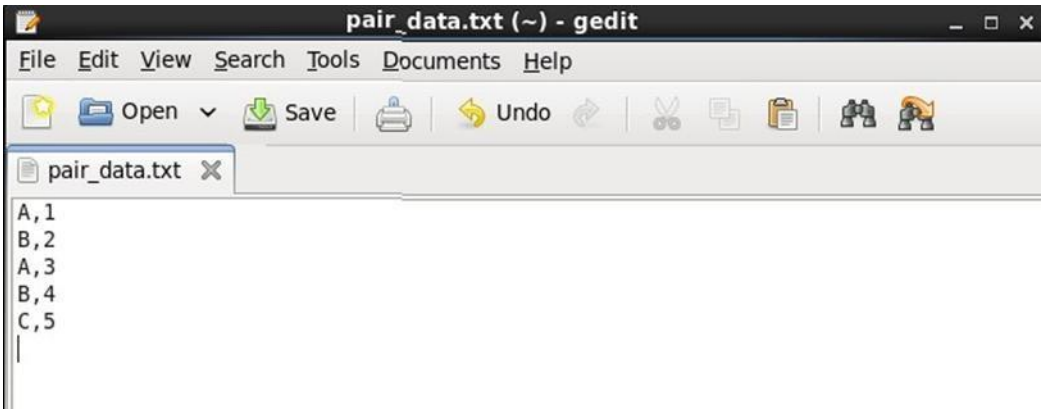
```
scala> rdd.saveAsTextFile("hdfs:///user/cloudera/rdd_output")
25/11/06 00:54:40 INFO deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id
25/11/06 00:54:40 INFO deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id
25/11/06 00:54:40 INFO deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap
25/11/06 00:54:40 INFO deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition
25/11/06 00:54:40 INFO deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id
25/11/06 00:54:40 INFO FileOutputCommitter: File Output Committer Algorithm version is 1
25/11/06 00:54:40 INFO SparkContext: Starting job: saveAsTextFile at <console>:26
25/11/06 00:54:40 INFO DAGScheduler: Got job 6 (saveAsTextFile at <console>:26) with 2 output partitions (allowLocal=false)
25/11/06 00:54:40 INFO DAGScheduler: Final stage: Stage 6(saveAsTextFile at <console>:26)
25/11/06 00:54:40 INFO DAGScheduler: Parents of final stage: List()
25/11/06 00:54:40 INFO DAGScheduler: Missing parents: List()
25/11/06 00:54:40 INFO DAGScheduler: Submitting Stage 6 (MapPartitionsRDD[5] at saveAsTextFile at <console>:26), which has no missing parents
25/11/06 00:54:40 INFO MemoryStore: ensureFreeSpace(136904) called with curMem=310192, maxMem=278302556
25/11/06 00:54:40 INFO MemoryStore: Block broadcast_7 stored as values in memory (estimated size 133.7 KB, free 265.0 MB)
25/11/06 00:54:40 INFO MemoryStore: ensureFreeSpace(47369) called with curMem=447096, maxMem=278302556
25/11/06 00:54:40 INFO MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 46.3 KB, free 264.9 MB)
25/11/06 00:54:40 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on quickstart.cloudera:54520 (size: 46.3 KB, free: 265.3 MB)
25/11/06 00:54:40 INFO BlockManagerMaster: Updated info of block broadcast_7_piece0
25/11/06 00:54:40 INFO SparkContext: Created broadcast 7 from broadcast at DAGScheduler.scala:839
25/11/06 00:54:40 INFO DAGScheduler: Submitting 2 missing tasks from Stage 6 (MapPartitionsRDD[5] at saveAsTextFile at <console>:26)
25/11/06 00:54:40 INFO YarnScheduler: Adding task set 6.0 with 2 tasks
25/11/06 00:54:40 INFO TaskSetManager: Starting task 0.0 in stage 6.0 (TID 12, quickstart.cloudera, PROCESS_LOCAL, 1162 bytes)
25/11/06 00:54:40 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on quickstart.cloudera:44681 (size: 46.3 KB, free: 530.2 MB)
25/11/06 00:54:43 INFO TaskSetManager: Starting task 1.0 in stage 6.0 (TID 13, quickstart.cloudera, PROCESS_LOCAL, 1163 bytes)
25/11/06 00:54:43 INFO TaskSetManager: Finished task 0.0 in stage 6.0 (TID 12) in 2565 ms on quickstart.cloudera (1/2)
25/11/06 00:54:43 INFO DAGScheduler: Stage 6 (saveAsTextFile at <console>:26) finished in 3.222 s
25/11/06 00:54:43 INFO DAGScheduler: Job 6 finished: saveAsTextFile at <console>:26, took 3.430018 s
25/11/06 00:54:43 INFO TaskSetManager: Finished task 1.0 in stage 6.0 (TID 13) in 661 ms on quickstart.cloudera (2/2)
25/11/06 00:54:43 INFO YarnScheduler: Removed TaskSet 6.0, whose tasks have all completed, from pool

scala> exit;
warning: there were 1 deprecation warning(s); re-run with -deprecation for details
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/zookeeper/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/jars/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/parquet/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera/rdd_output
Found 3 items
-rw-r--r--   1 cloudera cloudera          0 2025-11-06 00:54 /user/cloudera/rdd_output/_SUCCESS
-rw-r--r--   1 cloudera cloudera          4 2025-11-06 00:54 /user/cloudera/rdd_output/part-00000
-rw-r--r--   1 cloudera cloudera          6 2025-11-06 00:54 /user/cloudera/rdd_output/part-00001
```

Cache/Persist RDD:

```
scala> val rdd = sc.parallelize(Seq(1, 2, 3, 4, 5))
rdd: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[0] at parallelize at <console>:21

scala> rdd.cache()
res2: rdd.type = ParallelCollectionRDD[0] at parallelize at <console>:21

scala> rdd.count()
25/11/06 00:58:46 INFO SparkContext: Starting job: count at <console>:24
25/11/06 00:58:46 INFO DAGScheduler: Got job 0 (count at <console>:24) with 2 output partitions (allowLocal=false)
25/11/06 00:58:46 INFO DAGScheduler: Final stage: Stage 0(count at <console>:24)
25/11/06 00:58:46 INFO DAGScheduler: Parents of final stage: List()
25/11/06 00:58:46 INFO DAGScheduler: Missing parents: List()
25/11/06 00:58:46 INFO DAGScheduler: Submitting Stage 0 (ParallelCollectionRDD[0] at parallelize at <console>:21), which has no missing parents
25/11/06 00:58:47 INFO MemoryStore: ensureFreeSpace(1088) called with curMem=0, maxMem=278302556
25/11/06 00:58:47 INFO MemoryStore: Block broadcast_0 stored as values in memory (estimated size 1088.0 B, free 265.4 MB)
25/11/06 00:58:48 INFO MemoryStore: ensureFreeSpace(785) called with curMem=1088, maxMem=278302556
25/11/06 00:58:48 INFO MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 785.0 B, free 265.4 MB)
25/11/06 00:58:48 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on quickstart.cloudera:54996 (size: 785.0 B, free: 265.4 MB)
25/11/06 00:58:48 INFO BlockManagerMaster: Updated info of block broadcast_0_piece0
25/11/06 00:58:48 INFO SparkContext: Created broadcast 0 from broadcast at DAGScheduler.scala:839
25/11/06 00:58:48 INFO DAGScheduler: Submitting 2 missing tasks from Stage 0 (ParallelCollectionRDD[0] at parallelize at <console>:21)
25/11/06 00:58:48 INFO YarnScheduler: Adding task set 0.0 with 2 tasks
25/11/06 00:58:48 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, quickstart.cloudera, PROCESS_LOCAL, 1162 bytes)
25/11/06 00:58:50 INFO BlockManagerInfo: Added broadcast_0_piece0 in memory on quickstart.cloudera:57065 (size: 785.0 B, free: 530.3 MB)
25/11/06 00:58:51 INFO BlockManagerInfo: Added rdd_0_0 in memory on quickstart.cloudera:57065 (size: 72.0 B, free: 530.3 MB)
25/11/06 00:58:51 INFO TaskSetManager: Starting task 1.0 in stage 0.0 (TID 1, quickstart.cloudera, PROCESS_LOCAL, 1163 bytes)
25/11/06 00:58:51 INFO BlockManagerInfo: Added rdd_0_1 in memory on quickstart.cloudera:57065 (size: 104.0 B, free: 530.3 MB)
25/11/06 00:58:51 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 3067 ms on quickstart.cloudera (1/2)
25/11/06 00:58:51 INFO TaskSetManager: Finished task 1.0 in stage 0.0 (TID 1) in 180 ms on quickstart.cloudera (2/2)
25/11/06 00:58:51 INFO DAGScheduler: Stage 0 (count at <console>:24) finished in 3.170 s
25/11/06 00:58:51 INFO YarnScheduler: Removed TaskSet 0.0, whose tasks have all completed, from pool
25/11/06 00:58:51 INFO DAGScheduler: Job 0 finished: count at <console>:24, took 5.095243 s
res3: Long = 5
```

Create Pair RDD:

BIGDATA




```
scala> val pairRdd = sc.textFile("file:/home/cloudera/pair_data.txt").filter(line => line.contains(",")).map(line => { val parts = line.split(","); (parts(0), parts(
1).toInt) })
25/11/06 01:04:11 INFO MemoryStore: ensureFreeSpace(283092) called with curMem=311728, maxMem=278302556
25/11/06 01:04:11 INFO MemoryStore: Block broadcast_3 stored as values in memory (estimated size 276.5 KB, free 264.8 MB)
25/11/06 01:04:11 INFO BlockManager: Removing broadcast 2
25/11/06 01:04:11 INFO MemoryStore: ensureFreeSpace(22296) called with curMem=594820, maxMem=278302556
25/11/06 01:04:11 INFO MemoryStore: Block broadcast_3_piece0 stored as bytes in memory (estimated size 21.8 KB, free 264.8 MB)
25/11/06 01:04:11 INFO BlockManager: Removing block broadcast_2
25/11/06 01:04:11 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on quickstart.cloudera:54996 (size: 21.8 KB, free: 265.4 MB)
25/11/06 01:04:11 INFO BlockManagerMaster: Updated info of block broadcast_3_piece0
25/11/06 01:04:11 INFO SparkContext: Created broadcast 3 from textFile at <console>:21
25/11/06 01:04:11 INFO MemoryStore: Block broadcast_2 of size 2816 dropped from memory (free 277688256)
25/11/06 01:04:11 INFO BlockManager: Removing block broadcast_2_piece0
25/11/06 01:04:11 INFO MemoryStore: Block broadcast_2_piece0 of size 1723 dropped from memory (free 277689979)
25/11/06 01:04:11 INFO BlockManagerInfo: Removed broadcast_2_piece0 on quickstart.cloudera:54996 in memory (size: 1723.0 B, free: 265.4 MB)
25/11/06 01:04:11 INFO BlockManagerMaster: Updated info of block broadcast_2_piece0
25/11/06 01:04:11 INFO BlockManagerInfo: Removed broadcast_2_piece0 on quickstart.cloudera:57065 in memory (size: 1723.0 B, free: 530.3 MB)
25/11/06 01:04:11 INFO ContextCleaner: Cleaned broadcast 2
pairRdd: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[7] at map at <console>:21

scala> pairRdd.collect()
25/11/06 01:04:23 INFO FileInputFormat: Total input paths to process : 1
25/11/06 01:04:23 INFO SparkContext: Starting job: collect at <console>:24
25/11/06 01:04:23 INFO DAGScheduler: Got job 2 (collect at <console>:24) with 2 output partitions (allowLocal=false)
25/11/06 01:04:23 INFO DAGScheduler: Final stage: Stage 2(collect at <console>:24)
25/11/06 01:04:23 INFO DAGScheduler: Parents of final stage: List()
25/11/06 01:04:23 INFO DAGScheduler: Missing parents: List()
25/11/06 01:04:23 INFO DAGScheduler: Submitting Stage 2 (MapPartitionsRDD[7] at map at <console>:21), which has no missing parents
25/11/06 01:04:23 INFO MemoryStore: ensureFreeSpace(3024) called with curMem=612577, maxMem=278302556
25/11/06 01:04:23 INFO MemoryStore: Block broadcast_4 stored as values in memory (estimated size 3.0 KB, free 264.8 MB)
25/11/06 01:04:23 INFO MemoryStore: ensureFreeSpace(1792) called with curMem=615601, maxMem=278302556
25/11/06 01:04:23 INFO MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (estimated size 1792.0 B, free 264.8 MB)
25/11/06 01:04:23 INFO BlockManagerInfo: Added broadcast_4_piece0 in memory on quickstart.cloudera:54996 (size: 1792.0 B, free: 265.4 MB)
25/11/06 01:04:23 INFO BlockManagerMaster: Updated info of block broadcast_4_piece0
25/11/06 01:04:23 INFO SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:839
25/11/06 01:04:23 INFO DAGScheduler: Submitting 2 missing tasks from Stage 2 (MapPartitionsRDD[7] at map at <console>:21)
25/11/06 01:04:23 INFO YarnScheduler: Adding task set 2.0 with 2 tasks
25/11/06 01:04:23 INFO TaskSetManager: Starting task 0.0 in stage 2.0 (TID 7, quickstart.cloudera, PROCESS_LOCAL, 1297 bytes)
25/11/06 01:04:23 INFO BlockManagerInfo: Added broadcast_4_piece0 in memory on quickstart.cloudera:57065 (size: 1792.0 B, free: 530.3 MB)
25/11/06 01:04:23 INFO BlockManagerInfo: Added broadcast_3_piece0 in memory on quickstart.cloudera:57065 (size: 21.8 KB, free: 530.2 MB)
25/11/06 01:04:23 INFO TaskSetManager: Starting task 1.0 in stage 2.0 (TID 8, quickstart.cloudera, PROCESS_LOCAL, 1297 bytes)
25/11/06 01:04:23 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 7) in 338 ms on quickstart.cloudera (1/2)
25/11/06 01:04:23 INFO DAGScheduler: Stage 2 (collect at <console>:24) finished in 0.376 s
25/11/06 01:04:23 INFO DAGScheduler: Job 2 finished: collect at <console>:24, took 0.406607 s
25/11/06 01:04:23 INFO TaskSetManager: Finished task 1.0 in stage 2.0 (TID 8) in 36 ms on quickstart.cloudera (2/2)
25/11/06 01:04:23 INFO YarnScheduler: Removed TaskSet 2.0, whose tasks have all completed, from pool
res5: Array[(String, Int)] = Array((A,1), (B,2), (A,3), (B,4), (C,5))
```

reduceByKey:

```
scala> val reduced = pairRdd.reduceByKey((a, b) => a + b)
25/11/06 01:05:54 INFO BlockManager: Removing broadcast 4
25/11/06 01:05:54 INFO BlockManager: Removing block broadcast_4
25/11/06 01:05:54 INFO MemoryStore: Block broadcast_4 of size 3024 dropped from memory (free 277688187)
25/11/06 01:05:54 INFO BlockManager: Removing block broadcast_4_piece0
25/11/06 01:05:54 INFO MemoryStore: Block broadcast_4_piece0 of size 1792 dropped from memory (free 277689979)
25/11/06 01:05:54 INFO BlockManagerInfo: Removed broadcast_4_piece0 on quickstart.cloudera:54996 in memory (size: 1792.0 B, free: 265.4 MB)
25/11/06 01:05:54 INFO BlockManagerMaster: Updated info of block broadcast_4_piece0
25/11/06 01:05:54 INFO BlockManagerInfo: Removed broadcast_4_piece0 on quickstart.cloudera:57065 in memory (size: 1792.0 B, free: 530.2 MB)
25/11/06 01:05:55 INFO ContextCleaner: Cleaned broadcast 4
reduced: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[8] at reduceByKey at <console>:23
```

```
scala> reduced.collect()
25/11/06 01:06:02 INFO SparkContext: Starting job: collect at <console>:26
25/11/06 01:06:02 INFO DAGScheduler: Registering RDD 7 (map at <console>:21)
25/11/06 01:06:02 INFO DAGScheduler: Got job 3 (collect at <console>:26) with 2 output partitions (allowLocal=false)
25/11/06 01:06:02 INFO DAGScheduler: Final stage: Stage 4(collect at <console>:26)
25/11/06 01:06:02 INFO DAGScheduler: Parents of final stage: List(Stage 3)
25/11/06 01:06:02 INFO DAGScheduler: Missing parents: List(Stage 3)
25/11/06 01:06:02 INFO DAGScheduler: Submitting Stage 3 (MapPartitionsRDD[7] at map at <console>:21), which has no missing parents
25/11/06 01:06:02 INFO MemoryStore: ensureFreeSpace(3752) called with curMem=612577, maxMem=278302556
25/11/06 01:06:02 INFO MemoryStore: Block broadcast_5 stored as values in memory (estimated size 3.7 KB, free 264.8 MB)
25/11/06 01:06:02 INFO MemoryStore: ensureFreeSpace(2181) called with curMem=616329, maxMem=278302556
25/11/06 01:06:02 INFO MemoryStore: Block broadcast_5_piece0 stored as bytes in memory (estimated size 2.1 KB, free 264.8 MB)
25/11/06 01:06:02 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on quickstart.cloudera:54996 (size: 2.1 KB, free: 265.4 MB)
25/11/06 01:06:02 INFO BlockManagerMaster: Updated info of block broadcast_5_piece0
25/11/06 01:06:02 INFO SparkContext: Created broadcast 5 from broadcast at DAGScheduler.scala:839
25/11/06 01:06:02 INFO DAGScheduler: Submitting 2 missing tasks from Stage 3 (MapPartitionsRDD[7] at map at <console>:21)
25/11/06 01:06:02 INFO YarnScheduler: Adding task set 3.0 with 2 tasks
25/11/06 01:06:02 INFO TaskSetManager: Starting task 0.0 in stage 3.0 (TID 9, quickstart.cloudera, PROCESS_LOCAL, 1286 bytes)
25/11/06 01:06:02 INFO BlockManagerInfo: Added broadcast_5_piece0 in memory on quickstart.cloudera:57065 (size: 2.1 KB, free: 530.2 MB)
25/11/06 01:06:02 INFO TaskSetManager: Starting task 1.0 in stage 3.0 (TID 10, quickstart.cloudera, PROCESS_LOCAL, 1286 bytes)
25/11/06 01:06:02 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 9) in 314 ms on quickstart.cloudera (1/2)
25/11/06 01:06:02 INFO DAGScheduler: Stage 3 (map at <console>:21) finished in 0.441 s
25/11/06 01:06:02 INFO DAGScheduler: looking for newly runnable stages
25/11/06 01:06:02 INFO DAGScheduler: running: Set()
25/11/06 01:06:02 INFO DAGScheduler: waiting: Set(Stage 4)
25/11/06 01:06:02 INFO DAGScheduler: failed: Set()
25/11/06 01:06:02 INFO TaskSetManager: Finished task 1.0 in stage 3.0 (TID 10) in 129 ms on quickstart.cloudera (2/2)
25/11/06 01:06:02 INFO YarnScheduler: Removed TaskSet 3.0, whose tasks have all completed, from pool
25/11/06 01:06:02 INFO DAGScheduler: Missing parents for Stage 4: List()
25/11/06 01:06:02 INFO DAGScheduler: Submitting Stage 4 (ShuffledRDD[8] at reduceByKey at <console>:23), which is now runnable
25/11/06 01:06:02 INFO MemoryStore: ensureFreeSpace(2128) called with curMem=618510, maxMem=278302556
25/11/06 01:06:02 INFO MemoryStore: Block broadcast_6 stored as values in memory (estimated size 2.1 KB, free 264.8 MB)
25/11/06 01:06:02 INFO MemoryStore: ensureFreeSpace(1310) called with curMem=620638, maxMem=278302556
25/11/06 01:06:02 INFO MemoryStore: Block broadcast_6_piece0 stored as bytes in memory (estimated size 1310.0 B, free 264.8 MB)
25/11/06 01:06:02 INFO BlockManagerMaster: Updated info of block broadcast_6_piece0
25/11/06 01:06:02 INFO SparkContext: Created broadcast 6 from broadcast at DAGScheduler.scala:839
25/11/06 01:06:02 INFO DAGScheduler: Submitting 2 missing tasks from Stage 4 (ShuffledRDD[8] at reduceByKey at <console>:23)
25/11/06 01:06:02 INFO YarnScheduler: Adding task set 4.0 with 2 tasks
25/11/06 01:06:02 INFO TaskSetManager: Starting task 0.0 in stage 4.0 (TID 11, quickstart.cloudera, PROCESS_LOCAL, 1056 bytes)
25/11/06 01:06:02 INFO BlockManagerInfo: Added broadcast_6_piece0 in memory on quickstart.cloudera:57065 (size: 1310.0 B, free: 530.2 MB)
25/11/06 01:06:02 INFO MapOutputTrackerMasterActor: Asked to send map output locations for shuffle 0 to sparkExecutor@quickstart.cloudera:40965
25/11/06 01:06:02 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 0 is 157 bytes
25/11/06 01:06:03 INFO TaskSetManager: Starting task 1.0 in stage 4.0 (TID 12, quickstart.cloudera, PROCESS_LOCAL, 1056 bytes)
25/11/06 01:06:03 INFO TaskSetManager: Finished task 0.0 in stage 4.0 (TID 11) in 438 ms on quickstart.cloudera (1/2)
25/11/06 01:06:03 INFO DAGScheduler: Stage 4 (collect at <console>:26) finished in 0.547 s
25/11/06 01:06:03 INFO DAGScheduler: Job 3 finished: collect at <console>:26, took 1.107679 s
res6: Array[(String, Int)] = Array((B,6), (A,4), (C,5))
```

groupByKey – with output viewing:

```
scala> val grouped = pairRdd.groupByKey()
grouped: org.apache.spark.rdd.RDD[(String, Iterable[Int])] = ShuffledRDD[9] at groupByKey at <console>:23

scala> grouped.collect().foreach(println)
25/11/06 01:07:28 INFO BlockManager: Removing broadcast 6
25/11/06 01:07:28 INFO BlockManager: Removing block broadcast_6_piece0
25/11/06 01:07:28 INFO MemoryStore: Block broadcast_6_piece0 of size 1310 dropped from memory (free 277681918)
25/11/06 01:07:28 INFO BlockManagerInfo: Removed broadcast_6_piece0 on quickstart.cloudera:54996 in memory (size: 1310.0 B, free: 265.4 MB)
25/11/06 01:07:28 INFO BlockManagerMaster: Updated info of block broadcast_6_piece0
25/11/06 01:07:28 INFO BlockManager: Removing block broadcast_6
25/11/06 01:07:28 INFO MemoryStore: Block broadcast_6 of size 2128 dropped from memory (free 277684046)
25/11/06 01:07:28 INFO BlockManagerInfo: Removed broadcast_6_piece0 on quickstart.cloudera:57065 in memory (size: 1310.0 B, free: 530.2 MB)
25/11/06 01:07:28 INFO ContextCleaner: Cleaned broadcast 6
25/11/06 01:07:28 INFO BlockManager: Removing broadcast 5
25/11/06 01:07:28 INFO BlockManager: Removing block broadcast_5_piece0
25/11/06 01:07:28 INFO MemoryStore: Block broadcast_5_piece0 of size 2181 dropped from memory (free 277686227)
25/11/06 01:07:28 INFO BlockManagerInfo: Removed broadcast_5_piece0 on quickstart.cloudera:54996 in memory (size: 2.1 KB, free: 265.4 MB)
25/11/06 01:07:28 INFO BlockManagerMaster: Updated info of block broadcast_5_piece0
25/11/06 01:07:28 INFO BlockManager: Removing block broadcast_5
25/11/06 01:07:28 INFO MemoryStore: Block broadcast_5 of size 3752 dropped from memory (free 277689979)
25/11/06 01:07:28 INFO BlockManagerInfo: Removed broadcast_5_piece0 on quickstart.cloudera:57065 in memory (size: 2.1 KB, free: 530.2 MB)
25/11/06 01:07:28 INFO ContextCleaner: Cleaned broadcast 5
25/11/06 01:07:28 INFO SparkContext: Starting job: collect at <console>:26
25/11/06 01:07:28 INFO DAGScheduler: Registering RDD 7 (map at <console>:21)
25/11/06 01:07:28 INFO DAGScheduler: Got job 4 (collect at <console>:26) with 2 output partitions (allowLocal=false)
25/11/06 01:07:28 INFO DAGScheduler: Final stage: Stage 6(collect at <console>:26)
25/11/06 01:07:28 INFO DAGScheduler: Parents of final stage: List(Stage 5)
25/11/06 01:07:28 INFO DAGScheduler: Missing parents: List(Stage 5)
25/11/06 01:07:28 INFO DAGScheduler: Submitting Stage 5 (MapPartitionsRDD[7] at map at <console>:21), which has no missing parents
25/11/06 01:07:28 INFO MemoryStore: ensureFreeSpace(4224) called with curMem=612577, maxMem=278302556
25/11/06 01:07:28 INFO MemoryStore: Block broadcast_7 stored as values in memory (estimated size 4.1 KB, free 264.8 MB)
25/11/06 01:07:28 INFO MemoryStore: ensureFreeSpace(2400) called with curMem=616801, maxMem=278302556
25/11/06 01:07:28 INFO MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 2.3 KB, free 264.8 MB)
25/11/06 01:07:28 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on quickstart.cloudera:54996 (size: 2.3 KB, free: 265.4 MB)
25/11/06 01:07:28 INFO BlockManagerMaster: Updated info of block broadcast_7_piece0
```

```
25/11/06 01:07:28 INFO SparkContext: Created broadcast 7 from broadcast at DAGScheduler.scala:839
25/11/06 01:07:28 INFO DAGScheduler: Submitting 2 missing tasks from Stage 5 (MapPartitionsRDD[7] at map at <console>:21)
25/11/06 01:07:28 INFO YarnScheduler: Adding task set 5.0 with 2 tasks
25/11/06 01:07:28 INFO TaskSetManager: Starting task 0.0 in stage 5.0 (TID 13, quickstart.cloudera, PROCESS_LOCAL, 1286 bytes)
25/11/06 01:07:28 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on quickstart.cloudera:57065 (size: 2.3 KB, free: 530.2 MB)
25/11/06 01:07:29 INFO TaskSetManager: Starting task 1.0 in stage 5.0 (TID 14, quickstart.cloudera, PROCESS_LOCAL, 1286 bytes)
25/11/06 01:07:29 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (TID 13) in 300 ms on quickstart.cloudera (1/2)
25/11/06 01:07:29 INFO DAGScheduler: Stage 5 (map at <console>:21) finished in 0.432 s
25/11/06 01:07:29 INFO DAGScheduler: looking for newly runnable stages
25/11/06 01:07:29 INFO DAGScheduler: running: Set()
25/11/06 01:07:29 INFO DAGScheduler: waiting: Set(Stage 6)
25/11/06 01:07:29 INFO DAGScheduler: failed: Set()
25/11/06 01:07:29 INFO TaskSetManager: Finished task 1.0 in stage 5.0 (TID 14) in 117 ms on quickstart.cloudera (2/2)
25/11/06 01:07:29 INFO YarnScheduler: Removed TaskSet 5.0, whose tasks have all completed, from pool
25/11/06 01:07:29 INFO DAGScheduler: Missing parents for Stage 6: List()
25/11/06 01:07:29 INFO DAGScheduler: Submitting Stage 6 (ShuffledRDD[9] at groupByKey at <console>:23), which is now runnable
25/11/06 01:07:29 INFO MemoryStore: ensureFreeSpace(4600) called with curMem=619201, maxMem=278302556
25/11/06 01:07:29 INFO MemoryStore: Block broadcast_8 stored as values in memory (estimated size 4.5 KB, free 264.8 MB)
25/11/06 01:07:29 INFO MemoryStore: ensureFreeSpace(2562) called with curMem=623801, maxMem=278302556
25/11/06 01:07:29 INFO MemoryStore: Block broadcast_8_piece0 stored as bytes in memory (estimated size 2.5 KB, free 264.8 MB)
25/11/06 01:07:29 INFO BlockManagerInfo: Added broadcast_8_piece0 in memory on quickstart.cloudera:54996 (size: 2.5 KB, free: 265.4 MB)
25/11/06 01:07:29 INFO BlockManagerMaster: Updated info of block broadcast_8_piece0
25/11/06 01:07:29 INFO SparkContext: Created broadcast 8 from broadcast at DAGScheduler.scala:839
25/11/06 01:07:29 INFO DAGScheduler: Submitting 2 missing tasks from Stage 6 (ShuffledRDD[9] at groupByKey at <console>:23)
25/11/06 01:07:29 INFO YarnScheduler: Adding task set 6.0 with 2 tasks
25/11/06 01:07:29 INFO TaskSetManager: Starting task 0.0 in stage 6.0 (TID 15, quickstart.cloudera, PROCESS_LOCAL, 1056 bytes)
25/11/06 01:07:29 INFO BlockManagerInfo: Added broadcast_8_piece0 in memory on quickstart.cloudera:57065 (size: 2.5 KB, free: 530.2 MB)
25/11/06 01:07:29 INFO MapOutputTrackerMasterActor: Asked to send map output locations for shuffle 1 to sparkExecutor@quickstart.cloudera:40965
25/11/06 01:07:29 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 1 is 159 bytes
25/11/06 01:07:29 INFO TaskSetManager: Starting task 1.0 in stage 6.0 (TID 16, quickstart.cloudera, PROCESS_LOCAL, 1056 bytes)
25/11/06 01:07:29 INFO TaskSetManager: Finished task 0.0 in stage 6.0 (TID 15) in 176 ms on quickstart.cloudera (1/2)
25/11/06 01:07:29 INFO DAGScheduler: Stage 6 (collect at <console>:26) finished in 0.298 s
25/11/06 01:07:29 INFO DAGScheduler: Job 4 finished: collect at <console>:26, took 0.803611 s
25/11/06 01:07:29 INFO TaskSetManager: Finished task 1.0 in stage 6.0 (TID 16) in 115 ms on quickstart.cloudera (2/2)
25/11/06 01:07:29 INFO YarnScheduler: Removed TaskSet 6.0, whose tasks have all completed, from pool
(B,CompactBuffer(2, 4))
(A,CompactBuffer(1, 3))
(C,CompactBuffer(5))
```

mapValues:

```
scala> val doubled = pairRdd.mapValues(x => x * 2)
doubled: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[10] at mapValues at <console>:23

scala> doubled.collect()
25/11/06 01:09:41 INFO SparkContext: Starting job: collect at <console>:26
25/11/06 01:09:41 INFO DAGScheduler: Got job 5 (collect at <console>:26) with 2 output partitions (allowLocal=false)
25/11/06 01:09:41 INFO DAGScheduler: Final stage: Stage 7(collect at <console>:26)
25/11/06 01:09:41 INFO DAGScheduler: Parents of final stage: List()
25/11/06 01:09:41 INFO DAGScheduler: Missing parents: List()
25/11/06 01:09:41 INFO DAGScheduler: Submitting Stage 7 (MapPartitionsRDD[10] at mapValues at <console>:23), which has no missing parents
25/11/06 01:09:41 INFO MemoryStore: ensureFreeSpace(3264) called with curMem=626363, maxMem=278302556
25/11/06 01:09:41 INFO MemoryStore: Block broadcast_9 stored as values in memory (estimated size 3.2 KB, free 264.8 MB)
25/11/06 01:09:41 INFO MemoryStore: ensureFreeSpace(1858) called with curMem=629627, maxMem=278302556
25/11/06 01:09:41 INFO MemoryStore: Block broadcast_9_piece0 stored as bytes in memory (estimated size 1858.0 B, free 264.8 MB)
25/11/06 01:09:41 INFO BlockManagerInfo: Added broadcast_9_piece0 in memory on quickstart.cloudera:54996 (size: 1858.0 B, free: 265.4 MB)
25/11/06 01:09:41 INFO BlockManagerMaster: Updated info of block broadcast_9_piece0
25/11/06 01:09:41 INFO SparkContext: Created broadcast 9 from broadcast at DAGScheduler.scala:839
25/11/06 01:09:41 INFO DAGScheduler: Submitting 2 missing tasks from Stage 7 (MapPartitionsRDD[10] at mapValues at <console>:23)
25/11/06 01:09:41 INFO YarnScheduler: Adding task set 7.0 with 2 tasks
25/11/06 01:09:41 INFO TaskSetManager: Starting task 0.0 in stage 7.0 (TID 17, quickstart.cloudera, PROCESS_LOCAL, 1297 bytes)
25/11/06 01:09:41 INFO BlockManagerInfo: Added broadcast_9_piece0 in memory on quickstart.cloudera:57065 (size: 1858.0 B, free: 530.2 MB)
25/11/06 01:09:41 INFO TaskSetManager: Starting task 1.0 in stage 7.0 (TID 18, quickstart.cloudera, PROCESS_LOCAL, 1297 bytes)
25/11/06 01:09:41 INFO TaskSetManager: Finished task 0.0 in stage 7.0 (TID 17) in 114 ms on quickstart.cloudera (1/2)
25/11/06 01:09:41 INFO TaskSetManager: Finished task 1.0 in stage 7.0 (TID 18) in 61 ms on quickstart.cloudera (2/2)
25/11/06 01:09:41 INFO YarnScheduler: Removed TaskSet 7.0, whose tasks have all completed, from pool
25/11/06 01:09:41 INFO DAGScheduler: Stage 7 (collect at <console>:26) finished in 0.186 s
25/11/06 01:09:41 INFO DAGScheduler: Job 5 finished: collect at <console>:26, took 0.221699 s
res8: Array[(String, Int)] = Array((A,2), (B,4), (A,6), (B,8), (C,10))
```

sortByKey:

BIGDATA

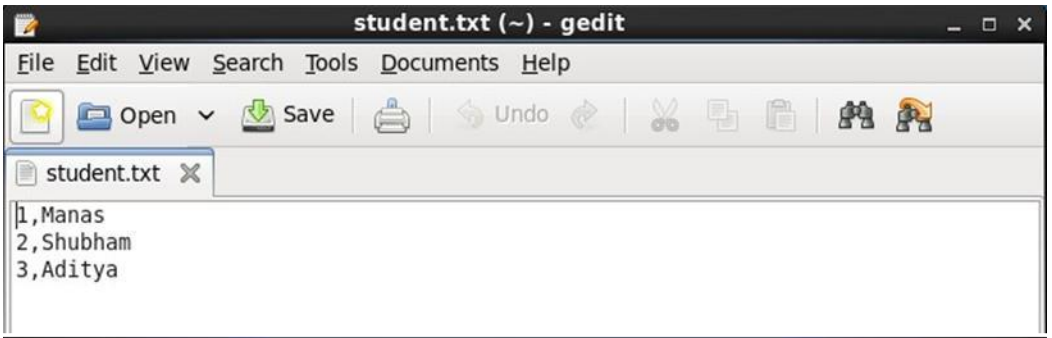
```
scala> val sorted = pairRdd.sortByKey()
25/11/06 01:10:18 INFO BlockManager: Removing broadcast 9
25/11/06 01:10:18 INFO BlockManager: Removing block broadcast_9
25/11/06 01:10:18 INFO MemoryStore: Block broadcast_9 of size 3264 dropped from memory (free 277674335)
25/11/06 01:10:18 INFO BlockManager: Removing block broadcast_9_piece0
25/11/06 01:10:18 INFO MemoryStore: Block broadcast_9_piece0 of size 1858 dropped from memory (free 277676193)
25/11/06 01:10:18 INFO BlockManagerInfo: Removed broadcast_9_piece0 on quickstart.cloudera:54996 in memory (size: 1858.0 B, free: 265.4 MB)
25/11/06 01:10:18 INFO BlockManagerMaster: Updated info of block broadcast_9_piece0
25/11/06 01:10:18 INFO BlockManagerInfo: Removed broadcast_9_piece0 on quickstart.cloudera:57065 in memory (size: 1858.0 B, free: 530.2 MB)
25/11/06 01:10:18 INFO ContextCleaner: Cleaned broadcast 9
25/11/06 01:10:18 INFO BlockManager: Removing broadcast 8
25/11/06 01:10:18 INFO BlockManager: Removing block broadcast_8
25/11/06 01:10:18 INFO MemoryStore: Block broadcast_8 of size 4600 dropped from memory (free 277680793)
25/11/06 01:10:18 INFO BlockManager: Removing block broadcast_8_piece0
25/11/06 01:10:18 INFO MemoryStore: Block broadcast_8_piece0 of size 2562 dropped from memory (free 277683355)
25/11/06 01:10:18 INFO BlockManagerInfo: Removed broadcast_8_piece0 on quickstart.cloudera:54996 in memory (size: 2.5 KB, free: 265.4 MB)
25/11/06 01:10:18 INFO BlockManagerMaster: Updated info of block broadcast_8_piece0
25/11/06 01:10:18 INFO BlockManagerInfo: Removed broadcast_8_piece0 on quickstart.cloudera:57065 in memory (size: 2.5 KB, free: 530.2 MB)
25/11/06 01:10:18 INFO ContextCleaner: Cleaned broadcast 8
25/11/06 01:10:18 INFO BlockManager: Removing broadcast 7
25/11/06 01:10:18 INFO BlockManager: Removing block broadcast_7
25/11/06 01:10:18 INFO MemoryStore: Block broadcast_7 of size 4224 dropped from memory (free 277687579)
25/11/06 01:10:18 INFO BlockManager: Removing block broadcast_7_piece0
25/11/06 01:10:18 INFO MemoryStore: Block broadcast_7_piece0 of size 2400 dropped from memory (free 277689979)
25/11/06 01:10:18 INFO BlockManagerInfo: Removed broadcast_7_piece0 on quickstart.cloudera:54996 in memory (size: 2.3 KB, free: 265.4 MB)
25/11/06 01:10:18 INFO BlockManagerMaster: Updated info of block broadcast_7_piece0
25/11/06 01:10:18 INFO BlockManagerInfo: Removed broadcast_7_piece0 on quickstart.cloudera:57065 in memory (size: 2.3 KB, free: 530.2 MB)
25/11/06 01:10:18 INFO ContextCleaner: Cleaned broadcast 7
25/11/06 01:10:18 INFO SparkContext: Starting job: sortByKey at <console>:23
25/11/06 01:10:18 INFO DAGScheduler: Got job 6 (sortByKey at <console>:23) with 2 output partitions (allowLocal=false)
25/11/06 01:10:18 INFO DAGScheduler: Final stage: Stage 8(sortByKey at <console>:23)
25/11/06 01:10:18 INFO DAGScheduler: Parents of final stage: List()
25/11/06 01:10:18 INFO DAGScheduler: Missing parents: List()
25/11/06 01:10:18 INFO DAGScheduler: Submitting Stage 8 (MapPartitionsRDD[12] at sortByKey at <console>:23), which has no missing parents
25/11/06 01:10:18 INFO MemoryStore: ensureFreeSpace(3472) called with curMem=612577, maxMem=278302556
25/11/06 01:10:18 INFO MemoryStore: Block broadcast_10 stored as values in memory (estimated size 3.4 KB, free 264.8 MB)
25/11/06 01:10:18 INFO MemoryStore: ensureFreeSpace(1968) called with curMem=616049, maxMem=278302556
25/11/06 01:10:18 INFO MemoryStore: Block broadcast_10_piece0 stored as bytes in memory (estimated size 1968.0 B, free 264.8 MB)
25/11/06 01:10:18 INFO BlockManagerInfo: Added broadcast_10_piece0 in memory on quickstart.cloudera:54996 (size: 1968.0 B, free: 265.4 MB)
25/11/06 01:10:18 INFO BlockManagerMaster: Updated info of block broadcast_10_piece0
25/11/06 01:10:18 INFO SparkContext: Created broadcast 10 from broadcast at DAGScheduler.scala:839
```

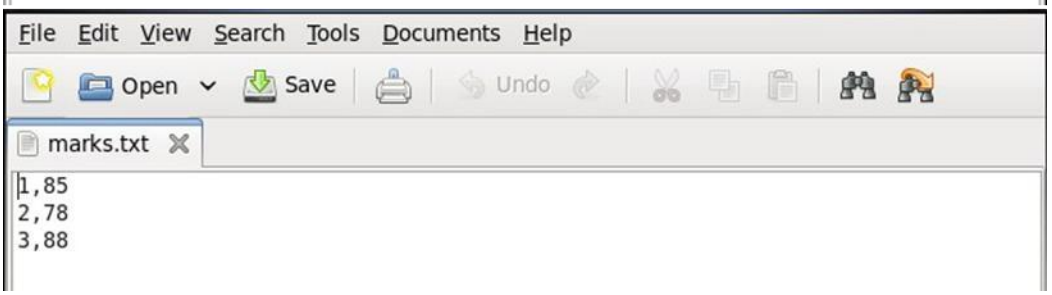
```
25/11/06 01:10:18 INFO DAGScheduler: Submitting 2 missing tasks from Stage 8 (MapPartitionsRDD[12] at sortByKey at <console>:23)
25/11/06 01:10:18 INFO YarnScheduler: Adding task set 8.0 with 2 tasks
25/11/06 01:10:18 INFO TaskSetManager: Starting task 0.0 in stage 8.0 (TID 19, quickstart.cloudera, PROCESS_LOCAL, 1297 bytes)
25/11/06 01:10:18 INFO BlockManagerInfo: Added broadcast_10_piece0 in memory on quickstart.cloudera:57065 (size: 1968.0 B, free: 530.2 MB)
25/11/06 01:10:18 INFO TaskSetManager: Starting task 1.0 in stage 8.0 (TID 20, quickstart.cloudera, PROCESS_LOCAL, 1297 bytes)
25/11/06 01:10:18 INFO TaskSetManager: Finished task 0.0 in stage 8.0 (TID 19) in 161 ms on quickstart.cloudera (1/2)
25/11/06 01:10:18 INFO DAGScheduler: Stage 8 (sortByKey at <console>:23) finished in 0.232 s
25/11/06 01:10:18 INFO TaskSetManager: Finished task 1.0 in stage 8.0 (TID 20) in 65 ms on quickstart.cloudera (2/2)
25/11/06 01:10:18 INFO YarnScheduler: Removed TaskSet 8.0, whose tasks have all completed, from pool
25/11/06 01:10:18 INFO DAGScheduler: Job 6 finished: sortByKey at <console>:23, took 0.271953 s
sorted: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[13] at sortByKey at <console>:23

scala> sorted.collect()
25/11/06 01:10:27 INFO SparkContext: Starting job: collect at <console>:26
25/11/06 01:10:27 INFO DAGScheduler: Registering RDD 7 (map at <console>:21)
25/11/06 01:10:27 INFO DAGScheduler: Got job 7 (collect at <console>:26) with 2 output partitions (allowLocal=false)
25/11/06 01:10:27 INFO DAGScheduler: Final stage: Stage 10(collect at <console>:26)
25/11/06 01:10:27 INFO DAGScheduler: Parents of final stage: List(Stage 9)
25/11/06 01:10:27 INFO DAGScheduler: Missing parents: List(Stage 9)
25/11/06 01:10:27 INFO DAGScheduler: Submitting Stage 9 (MapPartitionsRDD[7] at map at <console>:21), which has no missing parents
25/11/06 01:10:27 INFO MemoryStore: ensureFreeSpace(3992) called with curMem=618017, maxMem=278302556
25/11/06 01:10:27 INFO MemoryStore: Block broadcast_11 stored as values in memory (estimated size 3.9 KB, free 264.8 MB)
25/11/06 01:10:27 INFO MemoryStore: ensureFreeSpace(2283) called with curMem=622009, maxMem=278302556
25/11/06 01:10:27 INFO MemoryStore: Block broadcast_11_piece0 stored as bytes in memory (estimated size 2.2 KB, free 264.8 MB)
25/11/06 01:10:27 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on quickstart.cloudera:54996 (size: 2.2 KB, free: 265.4 MB)
25/11/06 01:10:27 INFO BlockManagerMaster: Updated info of block broadcast_11_piece0
25/11/06 01:10:27 INFO SparkContext: Created broadcast 11 from broadcast at DAGScheduler.scala:839
25/11/06 01:10:27 INFO DAGScheduler: Submitting 2 missing tasks from Stage 9 (MapPartitionsRDD[7] at map at <console>:21)
25/11/06 01:10:27 INFO YarnScheduler: Adding task set 9.0 with 2 tasks
25/11/06 01:10:27 INFO TaskSetManager: Starting task 0.0 in stage 9.0 (TID 21, quickstart.cloudera, PROCESS_LOCAL, 1286 bytes)
25/11/06 01:10:27 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on quickstart.cloudera:57065 (size: 2.2 KB, free: 530.2 MB)
25/11/06 01:10:28 INFO TaskSetManager: Starting task 1.0 in stage 9.0 (TID 22, quickstart.cloudera, PROCESS_LOCAL, 1286 bytes)
25/11/06 01:10:28 INFO TaskSetManager: Finished task 0.0 in stage 9.0 (TID 21) in 195 ms on quickstart.cloudera (1/2)
25/11/06 01:10:28 INFO DAGScheduler: Stage 9 (map at <console>:21) finished in 0.302 s
25/11/06 01:10:28 INFO DAGScheduler: looking for newly runnable stages
25/11/06 01:10:28 INFO DAGScheduler: running: Set()
25/11/06 01:10:28 INFO DAGScheduler: waiting: Set(Stage 10)
25/11/06 01:10:28 INFO DAGScheduler: failed: Set()
25/11/06 01:10:28 INFO DAGScheduler: Missing parents for Stage 10: List()
25/11/06 01:10:28 INFO DAGScheduler: Submitting Stage 10 (ShuffledRDD[13] at sortByKey at <console>:23), which is now runnable
25/11/06 01:10:28 INFO MemoryStore: ensureFreeSpace(2368) called with curMem=624292, maxMem=278302556
25/11/06 01:10:28 INFO MemoryStore: Block broadcast_12 stored as values in memory (estimated size 2.3 KB, free 264.8 MB)
25/11/06 01:10:28 INFO TaskSetManager: Finished task 1.0 in stage 9.0 (TID 22) in 103 ms on quickstart.cloudera (2/2)
25/11/06 01:10:28 INFO YarnScheduler: Removed TaskSet 9.0, whose tasks have all completed, from pool
```

```
25/11/06 01:10:28 INFO MemoryStore: ensureFreeSpace(1407) called with curMem=626660, maxMem=278302556
25/11/06 01:10:28 INFO MemoryStore: Block broadcast_12_piece0 stored as bytes in memory (estimated size 1407.0 B, free 264.8 MB)
25/11/06 01:10:28 INFO BlockManagerInfo: Added broadcast_12_piece0 in memory on quickstart.cloudera:54996 (size: 1407.0 B, free: 265.4 MB)
25/11/06 01:10:28 INFO BlockManagerMaster: Updated info of block broadcast_12_piece0
25/11/06 01:10:28 INFO SparkContext: Created broadcast 12 from broadcast at DAGScheduler.scala:839
25/11/06 01:10:28 INFO DAGScheduler: Submitting 2 missing tasks from Stage 10 (ShuffledRDD[13] at sortByKey at <console>:23)
25/11/06 01:10:28 INFO YarnScheduler: Adding task set 10.0 with 2 tasks
25/11/06 01:10:28 INFO TaskSetManager: Starting task 0.0 in stage 10.0 (TID 23, quickstart.cloudera, PROCESS_LOCAL, 1056 bytes)
25/11/06 01:10:28 INFO BlockManagerInfo: Added broadcast_12_piece0 in memory on quickstart.cloudera:57065 (size: 1407.0 B, free: 530.2 MB)
25/11/06 01:10:28 INFO MapOutputTrackerMasterActor: Asked to send map output locations for shuffle 2 to sparkExecutor@quickstart.cloudera:40965
25/11/06 01:10:28 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 2 is 160 bytes
25/11/06 01:10:28 INFO TaskSetManager: Starting task 1.0 in stage 10.0 (TID 24, quickstart.cloudera, PROCESS_LOCAL, 1056 bytes)
25/11/06 01:10:28 INFO TaskSetManager: Finished task 0.0 in stage 10.0 (TID 23) in 137 ms on quickstart.cloudera (1/2)
25/11/06 01:10:28 INFO DAGScheduler: Stage 10 (collect at <console>:26) finished in 0.239 s
25/11/06 01:10:28 INFO DAGScheduler: Job 7 finished: collect at <console>:26, took 0.642135 s
25/11/06 01:10:28 INFO TaskSetManager: Finished task 1.0 in stage 10.0 (TID 24) in 109 ms on quickstart.cloudera (2/2)
25/11/06 01:10:28 INFO YarnScheduler: Removed TaskSet 10.0, whose tasks have all completed, from pool
res9: Array[(String, Int)] = Array((A,1), (A,3), (B,2), (B,4), (C,5))
```

Join:

BIGDATA






```
scala> val studentRdd = sc.textFile("file:/home/cloudera/student.txt").map(x => { val p = x.split(","); (p(0).toInt, p(1)) })
25/11/06 01:14:36 INFO BlockManager: Removing broadcast 12
25/11/06 01:14:36 INFO BlockManager: Removing block broadcast_12
25/11/06 01:14:36 INFO MemoryStore: Block broadcast_12 of size 2368 dropped from memory (free 277676857)
25/11/06 01:14:36 INFO BlockManager: Removing block broadcast_12_piece0
25/11/06 01:14:36 INFO MemoryStore: Block broadcast_12_piece0 of size 1407 dropped from memory (free 277678264)
25/11/06 01:14:36 INFO BlockManagerInfo: Removed broadcast_12_piece0 on quickstart.cloudera:54996 in memory (size: 1407.0 B, free: 265.4 MB)
25/11/06 01:14:36 INFO BlockManagerMaster: Updated info of block broadcast_12_piece0
25/11/06 01:14:36 INFO BlockManagerInfo: Removed broadcast_12_piece0 on quickstart.cloudera:57065 in memory (size: 1407.0 B, free: 530.2 MB)
25/11/06 01:14:36 INFO ContextCleaner: Cleaned broadcast 12
25/11/06 01:14:36 INFO BlockManager: Removing broadcast 11
25/11/06 01:14:36 INFO BlockManager: Removing block broadcast_11
25/11/06 01:14:36 INFO MemoryStore: Block broadcast_11 of size 3992 dropped from memory (free 277682256)
25/11/06 01:14:36 INFO BlockManager: Removing block broadcast_11_piece0
25/11/06 01:14:36 INFO MemoryStore: Block broadcast_11_piece0 of size 2283 dropped from memory (free 277684539)
25/11/06 01:14:36 INFO BlockManagerInfo: Removed broadcast_11_piece0 on quickstart.cloudera:54996 in memory (size: 2.2 KB, free: 265.4 MB)
25/11/06 01:14:36 INFO BlockManagerMaster: Updated info of block broadcast_11_piece0
25/11/06 01:14:36 INFO BlockManagerInfo: Removed broadcast_11_piece0 on quickstart.cloudera:57065 in memory (size: 2.2 KB, free: 530.2 MB)
25/11/06 01:14:36 INFO ContextCleaner: Cleaned broadcast 11
25/11/06 01:14:36 INFO BlockManager: Removing broadcast 10
25/11/06 01:14:36 INFO BlockManager: Removing block broadcast_10
25/11/06 01:14:36 INFO MemoryStore: Block broadcast_10 of size 3472 dropped from memory (free 277688011)
25/11/06 01:14:36 INFO BlockManager: Removing block broadcast_10_piece0
25/11/06 01:14:36 INFO MemoryStore: Block broadcast_10_piece0 of size 1968 dropped from memory (free 277689979)
25/11/06 01:14:36 INFO BlockManagerInfo: Removed broadcast_10_piece0 on quickstart.cloudera:54996 in memory (size: 1968.0 B, free: 265.4 MB)
25/11/06 01:14:36 INFO BlockManagerMaster: Updated info of block broadcast_10_piece0
25/11/06 01:14:36 INFO MemoryStore: ensureFreeSpace(283092) called with curMem=612577, maxMem=278302556
25/11/06 01:14:36 INFO MemoryStore: Block broadcast_13 stored as values in memory (estimated size 276.5 KB, free 264.6 MB)
25/11/06 01:14:36 INFO BlockManagerInfo: Removed broadcast_10_piece0 on quickstart.cloudera:57065 in memory (size: 1968.0 B, free: 530.2 MB)
25/11/06 01:14:36 INFO ContextCleaner: Cleaned broadcast 10
25/11/06 01:14:36 INFO MemoryStore: ensureFreeSpace(22296) called with curMem=895669, maxMem=278302556
25/11/06 01:14:36 INFO MemoryStore: Block broadcast_13_piece0 stored as bytes in memory (estimated size 21.8 KB, free 264.5 MB)
25/11/06 01:14:36 INFO BlockManagerInfo: Added broadcast_13_piece0 in memory on quickstart.cloudera:54996 (size: 21.8 KB, free: 265.3 MB)
25/11/06 01:14:36 INFO BlockManagerMaster: Updated info of block broadcast_13_piece0
25/11/06 01:14:36 INFO SparkContext: Created broadcast 13 from textFile at <console>:21
studentRdd: org.apache.spark.rdd.RDD[(Int, String)] = MapPartitionsRDD[16] at map at <console>:21

scala> val marksRdd = sc.textFile("file:/home/cloudera/marks.txt").map(x => { val p = x.split(","); (p(0).toInt, p(1).toInt) })
25/11/06 01:14:51 INFO MemoryStore: ensureFreeSpace(283092) called with curMem=917965, maxMem=278302556
25/11/06 01:14:51 INFO MemoryStore: Block broadcast_14 stored as values in memory (estimated size 276.5 KB, free 264.3 MB)
25/11/06 01:14:51 INFO MemoryStore: ensureFreeSpace(22296) called with curMem=1201057, maxMem=278302556
25/11/06 01:14:51 INFO MemoryStore: Block broadcast_14_piece0 stored as bytes in memory (estimated size 21.8 KB, free 264.2 MB)
25/11/06 01:14:51 INFO BlockManagerInfo: Added broadcast_14_piece0 in memory on quickstart.cloudera:54996 (size: 21.8 KB, free: 265.3 MB)
25/11/06 01:14:51 INFO BlockManagerMaster: Updated info of block broadcast_14_piece0
25/11/06 01:14:51 INFO SparkContext: Created broadcast 14 from textFile at <console>:21
marksRdd: org.apache.spark.rdd.RDD[(Int, Int)] = MapPartitionsRDD[19] at map at <console>:21
```

BIGDATA

```
scala> val joined = studentRdd.join(marksRdd)
25/11/06 01:16:58 INFO FileInputFormat: Total input paths to process : 1
25/11/06 01:16:58 INFO FileInputFormat: Total input paths to process : 1
joined: org.apache.spark.rdd.RDD[(Int, (String, Int))] = MapPartitionsRDD[22] at join at <console>:25

scala> joined.collect()
25/11/06 01:17:08 INFO SparkContext: Starting job: collect at <console>:28
25/11/06 01:17:08 INFO DAGScheduler: Registering RDD 16 (map at <console>:21)
25/11/06 01:17:08 INFO DAGScheduler: Registering RDD 19 (map at <console>:21)
25/11/06 01:17:08 INFO DAGScheduler: Got job 8 (collect at <console>:28) with 3 output partitions (allowLocal=false)
25/11/06 01:17:08 INFO DAGScheduler: Final stage: Stage 13(collect at <console>:28)
25/11/06 01:17:08 INFO DAGScheduler: Parents of final stage: List(Stage 12, Stage 11)
25/11/06 01:17:08 INFO DAGScheduler: Missing parents: List(Stage 12, Stage 11)
25/11/06 01:17:08 INFO DAGScheduler: Submitting Stage 11 (MapPartitionsRDD[16] at map at <console>:21), which has no missing parents
25/11/06 01:17:08 INFO MemoryStore: ensureFreeSpace(3240) called with curMem=1223353, maxMem=278302556
25/11/06 01:17:08 INFO MemoryStore: Block broadcast_15 stored as values in memory (estimated size 3.2 KB, free 264.2 MB)
25/11/06 01:17:08 INFO MemoryStore: ensureFreeSpace(1981) called with curMem=1226593, maxMem=278302556
25/11/06 01:17:08 INFO MemoryStore: Block broadcast_15_piece0 stored as bytes in memory (estimated size 1981.0 B, free 264.2 MB)
25/11/06 01:17:08 INFO BlockManagerInfo: Added broadcast_15_piece0 in memory on quickstart.cloudera:54996 (size: 1981.0 B, free: 265.3 MB)
25/11/06 01:17:08 INFO BlockManagerMaster: Updated info of block broadcast_15_piece0
25/11/06 01:17:08 INFO SparkContext: Created broadcast 15 from broadcast at DAGScheduler.scala:839
25/11/06 01:17:08 INFO DAGScheduler: Submitting 2 missing tasks from Stage 11 (MapPartitionsRDD[16] at map at <console>:21)
25/11/06 01:17:08 INFO YarnScheduler: Adding task set 11.0 with 2 tasks
25/11/06 01:17:08 INFO TaskSetManager: Starting task 0.0 in stage 11.0 (TID 25, quickstart.cloudera, PROCESS_LOCAL, 1284 bytes)
25/11/06 01:17:08 INFO DAGScheduler: Submitting Stage 12 (MapPartitionsRDD[19] at map at <console>:21), which has no missing parents
25/11/06 01:17:08 INFO MemoryStore: ensureFreeSpace(3232) called with curMem=1228574, maxMem=278302556
25/11/06 01:17:08 INFO MemoryStore: Block broadcast_16 stored as values in memory (estimated size 3.2 KB, free 264.2 MB)
25/11/06 01:17:08 INFO BlockManagerInfo: Added broadcast_15_piece0 in memory on quickstart.cloudera:57065 (size: 1981.0 B, free: 530.2 MB)
25/11/06 01:17:08 INFO MemoryStore: ensureFreeSpace(1981) called with curMem=1231806, maxMem=278302556
25/11/06 01:17:08 INFO MemoryStore: Block broadcast_16_piece0 stored as bytes in memory (estimated size 1981.0 B, free 264.2 MB)
25/11/06 01:17:08 INFO BlockManagerInfo: Added broadcast_16_piece0 in memory on quickstart.cloudera:54996 (size: 1981.0 B, free: 265.3 MB)
25/11/06 01:17:08 INFO BlockManagerMaster: Updated info of block broadcast_16_piece0
25/11/06 01:17:08 INFO SparkContext: Created broadcast 16 from broadcast at DAGScheduler.scala:839
25/11/06 01:17:08 INFO DAGScheduler: Submitting 3 missing tasks from Stage 12 (MapPartitionsRDD[19] at map at <console>:21)
25/11/06 01:17:08 INFO MemoryStore: Block broadcast_16_piece0 stored as bytes in memory (estimated size 1981.0 B, free 264.2 MB)
25/11/06 01:17:08 INFO BlockManagerInfo: Added broadcast_16_piece0 in memory on quickstart.cloudera:54996 (size: 1981.0 B, free: 265.3 MB)
25/11/06 01:17:08 INFO BlockManagerMaster: Updated info of block broadcast_16_piece0
25/11/06 01:17:08 INFO SparkContext: Created broadcast 16 from broadcast at DAGScheduler.scala:839
25/11/06 01:17:08 INFO DAGScheduler: Submitting 3 missing tasks from Stage 12 (MapPartitionsRDD[19] at map at <console>:21)
25/11/06 01:17:08 INFO YarnScheduler: Adding task set 12.0 with 3 tasks
25/11/06 01:17:08 INFO BlockManagerInfo: Added broadcast_13_piece0 in memory on quickstart.cloudera:57065 (size: 21.8 KB, free: 530.2 MB)
25/11/06 01:17:09 INFO TaskSetManager: Starting task 1.0 in stage 11.0 (TID 26, quickstart.cloudera, PROCESS_LOCAL, 1284 bytes)
25/11/06 01:17:09 INFO TaskSetManager: Finished task 0.0 in stage 11.0 (TID 25) in 406 ms on quickstart.cloudera (1/2)
25/11/06 01:17:09 INFO TaskSetManager: Starting task 0.0 in stage 12.0 (TID 27, quickstart.cloudera, PROCESS_LOCAL, 1282 bytes)
25/11/06 01:17:09 INFO BlockManagerInfo: Added broadcast_16_piece0 in memory on quickstart.cloudera:57065 (size: 1981.0 B, free: 530.2 MB)
25/11/06 01:17:09 INFO TaskSetManager: Finished task 1.0 in stage 11.0 (TID 26) in 165 ms on quickstart.cloudera (2/2)
25/11/06 01:17:09 INFO DAGScheduler: Stage 11 (map at <console>:21) finished in 0.610 s
25/11/06 01:17:09 INFO DAGScheduler: looking for newly runnable stages
25/11/06 01:17:09 INFO DAGScheduler: running: Set(Stage 12)
25/11/06 01:17:09 INFO DAGScheduler: waiting: Set(Stage 13)
25/11/06 01:17:09 INFO DAGScheduler: failed: Set()
25/11/06 01:17:09 INFO YarnScheduler: Removed TaskSet 11.0, whose tasks have all completed, from pool
25/11/06 01:17:09 INFO DAGScheduler: Missing parents for Stage 13: List(Stage 12)
25/11/06 01:17:09 INFO BlockManagerInfo: Added broadcast_14_piece0 in memory on quickstart.cloudera:57065 (size: 21.8 KB, free: 530.2 MB)
25/11/06 01:17:09 INFO TaskSetManager: Starting task 1.0 in stage 12.0 (TID 28, quickstart.cloudera, PROCESS_LOCAL, 1282 bytes)
25/11/06 01:17:09 INFO TaskSetManager: Finished task 0.0 in stage 12.0 (TID 27) in 311 ms on quickstart.cloudera (1/3)
25/11/06 01:17:09 INFO TaskSetManager: Starting task 2.0 in stage 12.0 (TID 29, quickstart.cloudera, PROCESS_LOCAL, 1282 bytes)
25/11/06 01:17:09 INFO TaskSetManager: Finished task 1.0 in stage 12.0 (TID 28) in 98 ms on quickstart.cloudera (2/3)
25/11/06 01:17:09 INFO DAGScheduler: Stage 12 (map at <console>:21) finished in 0.932 s
25/11/06 01:17:09 INFO DAGScheduler: looking for newly runnable stages
25/11/06 01:17:09 INFO DAGScheduler: running: Set()
25/11/06 01:17:09 INFO DAGScheduler: waiting: Set(Stage 13)
25/11/06 01:17:09 INFO DAGScheduler: failed: Set()
25/11/06 01:17:09 INFO DAGScheduler: Missing parents for Stage 13: List()
25/11/06 01:17:09 INFO DAGScheduler: Submitting Stage 13 (MapPartitionsRDD[22] at join at <console>:25), which is now runnable
25/11/06 01:17:09 INFO TaskSetManager: Finished task 2.0 in stage 12.0 (TID 29) in 36 ms on quickstart.cloudera (3/3)
25/11/06 01:17:09 INFO YarnScheduler: Removed TaskSet 12.0, whose tasks have all completed, from pool
25/11/06 01:17:09 INFO MemoryStore: ensureFreeSpace(2568) called with curMem=1233787, maxMem=278302556
25/11/06 01:17:09 INFO MemoryStore: Block broadcast_17 stored as values in memory (estimated size 2.5 KB, free 264.2 MB)
25/11/06 01:17:09 INFO MemoryStore: ensureFreeSpace(1469) called with curMem=1236355, maxMem=278302556
25/11/06 01:17:09 INFO MemoryStore: Block broadcast_17_piece0 stored as bytes in memory (estimated size 1469.0 B, free 264.2 MB)
25/11/06 01:17:09 INFO BlockManagerInfo: Added broadcast_17_piece0 in memory on quickstart.cloudera:54996 (size: 1469.0 B, free: 265.3 MB)
25/11/06 01:17:09 INFO BlockManagerMaster: Updated info of block broadcast_17_piece0
25/11/06 01:17:09 INFO SparkContext: Created broadcast 17 from broadcast at DAGScheduler.scala:839
25/11/06 01:17:09 INFO DAGScheduler: Submitting 3 missing tasks from Stage 13 (MapPartitionsRDD[22] at join at <console>:25)
25/11/06 01:17:09 INFO YarnScheduler: Adding task set 13.0 with 3 tasks
25/11/06 01:17:09 INFO TaskSetManager: Starting task 0.0 in stage 13.0 (TID 30, quickstart.cloudera, PROCESS_LOCAL, 1902 bytes)
25/11/06 01:17:09 INFO BlockManagerInfo: Added broadcast_17_piece0 in memory on quickstart.cloudera:57065 (size: 1469.0 B, free: 530.2 MB)
25/11/06 01:17:09 INFO MapOutputTrackerMasterActor: Asked to send map output locations for shuffle 3 to sparkExecutor@quickstart.cloudera:40965
25/11/06 01:17:09 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 3 is 163 bytes
25/11/06 01:17:09 INFO MapOutputTrackerMasterActor: Asked to send map output locations for shuffle 4 to sparkExecutor@quickstart.cloudera:40965
25/11/06 01:17:09 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 4 is 166 bytes
25/11/06 01:17:09 INFO TaskSetManager: Finished task 0.0 in stage 13.0 (TID 30) in 242 ms on quickstart.cloudera (1/3)
25/11/06 01:17:09 INFO TaskSetManager: Starting task 1.0 in stage 13.0 (TID 31, quickstart.cloudera, PROCESS_LOCAL, 1902 bytes)
25/11/06 01:17:10 INFO TaskSetManager: Starting task 2.0 in stage 13.0 (TID 32, quickstart.cloudera, PROCESS_LOCAL, 1902 bytes)
25/11/06 01:17:10 INFO TaskSetManager: Finished task 1.0 in stage 13.0 (TID 31) in 128 ms on quickstart.cloudera (2/3)
25/11/06 01:17:10 INFO DAGScheduler: Stage 13 (collect at <console>:28) finished in 0.473 s
25/11/06 01:17:10 INFO DAGScheduler: Job 8 finished: collect at <console>:28, took 1.559666 s
res10: Array[(Int, (String, Int))] = Array((3,(Aditya,88)), (1,(Manas,85)), (2,(Shubham,78)))
```

Cogroup:

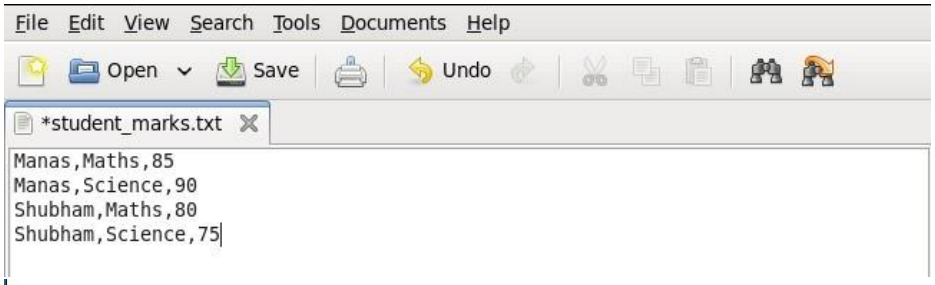
BIGDATA

```
scala> val cog = studentRdd.cogroup(marksRdd)
cog: org.apache.spark.rdd.RDD[(Int, (Iterable[String], Iterable[Int]))] = MapPartitionsRDD[24] at cogroup at <console>:25

scala> cog.collect().foreach(println)
25/11/06 01:18:49 INFO SparkContext: Starting job: collect at <console>:28
25/11/06 01:18:49 INFO DAGScheduler: Registering RDD 16 (map at <console>:21)
25/11/06 01:18:49 INFO DAGScheduler: Registering RDD 19 (map at <console>:21)
25/11/06 01:18:49 INFO DAGScheduler: Got job 9 (collect at <console>:28) with 3 output partitions (allowLocal=false)
25/11/06 01:18:49 INFO DAGScheduler: Final stage: Stage 16(collect at <console>:28)
25/11/06 01:18:49 INFO DAGScheduler: Parents of final stage: List(Stage 15, Stage 14)
25/11/06 01:18:49 INFO DAGScheduler: Missing parents: List(Stage 15, Stage 14)
25/11/06 01:18:49 INFO DAGScheduler: Submitting Stage 14 (MapPartitionsRDD[16] at map at <console>:21), which has no missing parents
25/11/06 01:18:49 INFO MemoryStore: ensureFreeSpace(3240) called with curMem=1237824, maxMem=278302556
25/11/06 01:18:49 INFO MemoryStore: Block broadcast_18 stored as values in memory (estimated size 3.2 KB, free 264.2 MB)
25/11/06 01:18:49 INFO MemoryStore: ensureFreeSpace(1983) called with curMem=1241064, maxMem=278302556
25/11/06 01:18:49 INFO MemoryStore: Block broadcast_18_piece0 stored as bytes in memory (estimated size 1983.0 B, free 264.2 MB)
25/11/06 01:18:49 INFO BlockManagerInfo: Added broadcast_18_piece0 in memory on quickstart.cloudera:54996 (size: 1983.0 B, free: 265.3 MB)
25/11/06 01:18:49 INFO BlockManagerMaster: Updated info of block broadcast_18_piece0
25/11/06 01:18:49 INFO SparkContext: Created broadcast 18 from broadcast at DAGScheduler.scala:839
25/11/06 01:18:49 INFO DAGScheduler: Submitting 2 missing tasks from Stage 14 (MapPartitionsRDD[16] at map at <console>:21)
25/11/06 01:18:49 INFO YarnScheduler: Adding task set 14.0 with 2 tasks
25/11/06 01:18:49 INFO TaskSetManager: Starting task 0.0 in stage 14.0 (TID 33, quickstart.cloudera, PROCESS_LOCAL, 1284 bytes)
25/11/06 01:18:49 INFO DAGScheduler: Submitting Stage 15 (MapPartitionsRDD[19] at map at <console>:21), which has no missing parents
25/11/06 01:18:49 INFO MemoryStore: ensureFreeSpace(3232) called with curMem=1243047, maxMem=278302556
25/11/06 01:18:49 INFO MemoryStore: Block broadcast_19 stored as values in memory (estimated size 3.2 KB, free 264.2 MB)
25/11/06 01:18:49 INFO BlockManager: Removing broadcast 17
25/11/06 01:18:49 INFO BlockManager: Removing block broadcast_17_piece0
25/11/06 01:18:49 INFO MemoryStore: Block broadcast_17_piece0 of size 1469 dropped from memory (free 277057746)
25/11/06 01:18:49 INFO BlockManagerInfo: Removed broadcast_17_piece0 on quickstart.cloudera:54996 in memory (size: 1469.0 B, free: 265.3 MB)
25/11/06 01:18:49 INFO BlockManagerMaster: Updated info of block broadcast_17_piece0
25/11/06 01:18:49 INFO BlockManager: Removing block broadcast_17
25/11/06 01:18:49 INFO MemoryStore: Block broadcast_17 of size 2568 dropped from memory (free 277060314)
25/11/06 01:18:49 INFO MemoryStore: ensureFreeSpace(1981) called with curMem=1242242, maxMem=278302556
25/11/06 01:18:49 INFO MemoryStore: Block broadcast_19_piece0 stored as bytes in memory (estimated size 1981.0 B, free 264.2 MB)
25/11/06 01:18:49 INFO BlockManagerInfo: Added broadcast_19_piece0 in memory on quickstart.cloudera:54996 (size: 1981.0 B, free: 265.3 MB)
25/11/06 01:18:49 INFO BlockManagerMaster: Updated info of block broadcast_19_piece0
25/11/06 01:18:49 INFO SparkContext: Created broadcast 19 from broadcast at DAGScheduler.scala:839
25/11/06 01:18:49 INFO DAGScheduler: Submitting 3 missing tasks from Stage 15 (MapPartitionsRDD[19] at map at <console>:21)
25/11/06 01:18:49 INFO YarnScheduler: Adding task set 15.0 with 3 tasks
25/11/06 01:18:49 INFO BlockManagerInfo: Added broadcast_18_piece0 in memory on quickstart.cloudera:57065 (size: 1983.0 B, free: 530.2 MB)
25/11/06 01:18:49 INFO BlockManagerInfo: Removed broadcast_17_piece0 on quickstart.cloudera:57065 in memory (size: 1469.0 B, free: 530.2 MB)
25/11/06 01:18:49 INFO ContextCleaner: Cleaned broadcast 17
25/11/06 01:18:49 INFO BlockManager: Removing broadcast 16
```

```
25/11/06 01:18:49 INFO BlockManager: Removing block broadcast_16_piece0
25/11/06 01:18:49 INFO MemoryStore: Block broadcast_16_piece0 of size 1981 dropped from memory (free 277060314)
25/11/06 01:18:49 INFO BlockManagerInfo: Removed broadcast_16_piece0 on quickstart.cloudera:54996 in memory (size: 1981.0 B, free: 265.3 MB)
25/11/06 01:18:49 INFO BlockManagerMaster: Updated info of block broadcast_16_piece0
25/11/06 01:18:49 INFO BlockManager: Removing block broadcast_16
25/11/06 01:18:49 INFO MemoryStore: Block broadcast_16 of size 3232 dropped from memory (free 277063546)
25/11/06 01:18:49 INFO BlockManagerInfo: Removed broadcast_16_piece0 on quickstart.cloudera:57065 in memory (size: 1981.0 B, free: 530.2 MB)
25/11/06 01:18:49 INFO ContextCleaner: Cleaned broadcast 16
25/11/06 01:18:50 INFO TaskSetManager: Starting task 1.0 in stage 14.0 (TID 34, quickstart.cloudera, PROCESS_LOCAL, 1284 bytes)
25/11/06 01:18:50 INFO TaskSetManager: Finished task 0.0 in stage 14.0 (TID 33) in 241 ms on quickstart.cloudera (1/2)
25/11/06 01:18:50 INFO TaskSetManager: Starting task 0.0 in stage 15.0 (TID 35, quickstart.cloudera, PROCESS_LOCAL, 1282 bytes)
25/11/06 01:18:50 INFO TaskSetManager: Finished task 1.0 in stage 14.0 (TID 34) in 129 ms on quickstart.cloudera (2/2)
25/11/06 01:18:50 INFO YarnScheduler: Removed TaskSet 14.0, whose tasks have all completed, from pool
25/11/06 01:18:50 INFO DAGScheduler: Stage 14 (map at <console>:21) finished in 0.372 s
25/11/06 01:18:50 INFO DAGScheduler: looking for newly runnable stages
25/11/06 01:18:50 INFO DAGScheduler: running: Set(Stage 15)
25/11/06 01:18:50 INFO DAGScheduler: waiting: Set(Stage 16)
25/11/06 01:18:50 INFO DAGScheduler: failed: Set()
25/11/06 01:18:50 INFO DAGScheduler: Missing parents for Stage 16: List(Stage 15)
25/11/06 01:18:50 INFO BlockManagerInfo: Added broadcast_19_piece0 in memory on quickstart.cloudera:57065 (size: 1981.0 B, free: 530.2 MB)
25/11/06 01:18:50 INFO TaskSetManager: Starting task 1.0 in stage 15.0 (TID 36, quickstart.cloudera, PROCESS_LOCAL, 1282 bytes)
25/11/06 01:18:50 INFO TaskSetManager: Finished task 0.0 in stage 15.0 (TID 35) in 173 ms on quickstart.cloudera (1/3)
25/11/06 01:18:50 INFO TaskSetManager: Starting task 2.0 in stage 15.0 (TID 37, quickstart.cloudera, PROCESS_LOCAL, 1282 bytes)
25/11/06 01:18:50 INFO TaskSetManager: Finished task 1.0 in stage 15.0 (TID 36) in 103 ms on quickstart.cloudera (2/3)
25/11/06 01:18:50 INFO DAGScheduler: Stage 15 (map at <console>:21) finished in 0.659 s
25/11/06 01:18:50 INFO DAGScheduler: looking for newly runnable stages
25/11/06 01:18:50 INFO DAGScheduler: running: Set()
25/11/06 01:18:50 INFO DAGScheduler: waiting: Set(Stage 16)
25/11/06 01:18:50 INFO DAGScheduler: failed: Set()
25/11/06 01:18:50 INFO DAGScheduler: Missing parents for Stage 16: List()
25/11/06 01:18:50 INFO DAGScheduler: Submitting Stage 16 (MapPartitionsRDD[24] at cogroup at <console>:25), which is now runnable
25/11/06 01:18:50 INFO MemoryStore: ensureFreeSpace(2336) called with curMem=1239010, maxMem=278302556
25/11/06 01:18:50 INFO MemoryStore: Block broadcast_20 stored as values in memory (estimated size 2.3 KB, free 264.2 MB)
25/11/06 01:18:50 INFO MemoryStore: ensureFreeSpace(1410) called with curMem=1241346, maxMem=278302556
25/11/06 01:18:50 INFO TaskSetManager: Finished task 2.0 in stage 15.0 (TID 37) in 78 ms on quickstart.cloudera (3/3)
25/11/06 01:18:50 INFO YarnScheduler: Removed TaskSet 15.0, whose tasks have all completed, from pool
25/11/06 01:18:50 INFO MemoryStore: Block broadcast_20_piece0 stored as bytes in memory (estimated size 1410.0 B, free 264.2 MB)
25/11/06 01:18:50 INFO BlockManagerInfo: Added broadcast_20_piece0 in memory on quickstart.cloudera:54996 (size: 1410.0 B, free: 265.3 MB)
25/11/06 01:18:50 INFO BlockManagerMaster: Updated info of block broadcast_20_piece0
25/11/06 01:18:50 INFO SparkContext: Created broadcast 20 from broadcast at DAGScheduler.scala:839
25/11/06 01:18:50 INFO DAGScheduler: Submitting 3 missing tasks from Stage 16 (MapPartitionsRDD[24] at cogroup at <console>:25)
25/11/06 01:18:50 INFO YarnScheduler: Adding task set 16.0 with 3 tasks
25/11/06 01:18:50 INFO TaskSetManager: Starting task 0.0 in stage 16.0 (TID 38, quickstart.cloudera, PROCESS_LOCAL, 1902 bytes)
25/11/06 01:18:50 INFO BlockManagerInfo: Added broadcast_20_piece0 in memory on quickstart.cloudera:57065 (size: 1410.0 B, free: 530.2 MB)
```

```
25/11/06 01:18:50 INFO MapOutputTrackerMasterActor: Asked to send map output locations for shuffle 5 to sparkExecutor@quickstart.cloudera:40965
25/11/06 01:18:50 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 5 is 163 bytes
25/11/06 01:18:50 INFO MapOutputTrackerMasterActor: Asked to send map output locations for shuffle 6 to sparkExecutor@quickstart.cloudera:40965
25/11/06 01:18:50 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 6 is 166 bytes
25/11/06 01:18:50 INFO TaskSetManager: Starting task 1.0 in stage 16.0 (TID 39, quickstart.cloudera, PROCESS_LOCAL, 1902 bytes)
25/11/06 01:18:50 INFO TaskSetManager: Finished task 0.0 in stage 16.0 (TID 38) in 227 ms on quickstart.cloudera (1/3)
25/11/06 01:18:50 INFO TaskSetManager: Starting task 2.0 in stage 16.0 (TID 40, quickstart.cloudera, PROCESS_LOCAL, 1902 bytes)
25/11/06 01:18:50 INFO TaskSetManager: Finished task 1.0 in stage 16.0 (TID 39) in 112 ms on quickstart.cloudera (2/3)
25/11/06 01:18:51 INFO TaskSetManager: Finished task 2.0 in stage 16.0 (TID 40) in 164 ms on quickstart.cloudera (3/3)
25/11/06 01:18:51 INFO DAGScheduler: Stage 16 (collect at <console>:28) finished in 0.496 s
25/11/06 01:18:51 INFO YarnScheduler: Removed TaskSet 16.0, whose tasks have all completed, from pool
25/11/06 01:18:51 INFO DAGScheduler: Job 9 finished: collect at <console>:28, took 1.336485 s
(3,(CompactBuffer(Aditya),CompactBuffer(88)))
(1,(CompactBuffer(Manas),CompactBuffer(85)))
(2,(CompactBuffer(Shubham),CompactBuffer(78)))
```

aggregateByKey:




```
scala> val marks = sc.textFile("file:/home/cloudera/student_marks.txt").map(x => { val p = x.split(","); (p(0), p(2).toInt) })
25/11/06 01:21:16 INFO MemoryStore: ensureFreeSpace(283092) called with curMem=1242756, maxMem=278302556
25/11/06 01:21:16 INFO MemoryStore: Block broadcast_21 stored as values in memory (estimated size 276.5 KB, free 264.0 MB)
25/11/06 01:21:16 INFO MemoryStore: ensureFreeSpace(22296) called with curMem=1525848, maxMem=278302556
25/11/06 01:21:16 INFO MemoryStore: Block broadcast_21_piece0 stored as bytes in memory (estimated size 21.8 KB, free 263.9 MB)
25/11/06 01:21:16 INFO BlockManagerInfo: Added broadcast_21_piece0 in memory on quickstart.cloudera:54996 (size: 21.8 KB, free: 265.3 MB)
25/11/06 01:21:16 INFO BlockManagerMaster: Updated info of block broadcast_21_piece0
25/11/06 01:21:16 INFO SparkContext: Created broadcast 21 from textFile at <console>:21
marks: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[27] at map at <console>:21

scala> val agg = marks.aggregateByKey(0)((acc, v) => acc + v, (acc1, acc2) => acc1 + acc2)
25/11/06 01:21:32 INFO FileInputFormat: Total input paths to process : 1
agg: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[28] at aggregateByKey at <console>:23
```

```
scala> agg.collect()
25/11/06 01:22:04 INFO BlockManager: Removing broadcast 20
25/11/06 01:22:04 INFO BlockManager: Removing block broadcast_20_piece0
25/11/06 01:22:04 INFO MemoryStore: Block broadcast_20_piece0 of size 1410 dropped from memory (free 276755822)
25/11/06 01:22:04 INFO BlockManagerInfo: Removed broadcast_20_piece0 on quickstart.cloudera:54996 in memory (size: 1410.0 B, free: 265.3 MB)
25/11/06 01:22:04 INFO BlockManagerMaster: Updated info of block broadcast_20_piece0
25/11/06 01:22:04 INFO BlockManager: Removing block broadcast_20
25/11/06 01:22:04 INFO MemoryStore: Block broadcast_20 of size 2336 dropped from memory (free 276758158)
25/11/06 01:22:04 INFO BlockManagerInfo: Removed broadcast_20_piece0 on quickstart.cloudera:57065 in memory (size: 1410.0 B, free: 530.2 MB)
25/11/06 01:22:04 INFO ContextCleaner: Cleaned broadcast 20
25/11/06 01:22:04 INFO BlockManager: Removing broadcast 19
25/11/06 01:22:04 INFO BlockManager: Removing block broadcast_19
25/11/06 01:22:04 INFO MemoryStore: Block broadcast_19 of size 3232 dropped from memory (free 276761390)
25/11/06 01:22:04 INFO BlockManager: Removing block broadcast_19_piece0
25/11/06 01:22:04 INFO MemoryStore: Block broadcast_19_piece0 of size 1981 dropped from memory (free 276763371)
25/11/06 01:22:04 INFO BlockManagerInfo: Removed broadcast_19_piece0 on quickstart.cloudera:54996 in memory (size: 1981.0 B, free: 265.3 MB)
25/11/06 01:22:04 INFO BlockManagerMaster: Updated info of block broadcast_19_piece0
25/11/06 01:22:04 INFO BlockManagerInfo: Removed broadcast_19_piece0 on quickstart.cloudera:57065 in memory (size: 1981.0 B, free: 530.2 MB)
25/11/06 01:22:04 INFO ContextCleaner: Cleaned broadcast 19
25/11/06 01:22:05 INFO SparkContext: Starting job: collect at <console>:26
25/11/06 01:22:05 INFO DAGScheduler: Registering RDD 27 (map at <console>:21)
25/11/06 01:22:05 INFO DAGScheduler: Got job 10 (collect at <console>:26) with 2 output partitions (allowLocal=false)
25/11/06 01:22:05 INFO DAGScheduler: Final stage: Stage 18(collect at <console>:26)
25/11/06 01:22:05 INFO DAGScheduler: Parents of final stage: List(Stage 17)
25/11/06 01:22:05 INFO DAGScheduler: Missing parents: List(Stage 17)
25/11/06 01:22:05 INFO DAGScheduler: Submitting Stage 17 (MapPartitionsRDD[27] at map at <console>:21), which has no missing parents
25/11/06 01:22:05 INFO MemoryStore: ensureFreeSpace(4464) called with curMem=1539185, maxMem=278302556
25/11/06 01:22:05 INFO MemoryStore: Block broadcast_22 stored as values in memory (estimated size 4.4 KB, free 263.9 MB)
25/11/06 01:22:05 INFO MemoryStore: ensureFreeSpace(2570) called with curMem=1543649, maxMem=278302556
25/11/06 01:22:05 INFO MemoryStore: Block broadcast_22_piece0 stored as bytes in memory (estimated size 2.5 KB, free 263.9 MB)
25/11/06 01:22:05 INFO BlockManagerInfo: Added broadcast_22_piece0 in memory on quickstart.cloudera:54996 (size: 2.5 KB, free: 265.3 MB)
25/11/06 01:22:05 INFO BlockManagerMaster: Updated info of block broadcast_22_piece0
25/11/06 01:22:05 INFO SparkContext: Created broadcast 22 from broadcast at DAGScheduler.scala:839
25/11/06 01:22:05 INFO DAGScheduler: Submitting 2 missing tasks from Stage 17 (MapPartitionsRDD[27] at map at <console>:21)
25/11/06 01:22:05 INFO YarnScheduler: Adding task set 17.0 with 2 tasks
25/11/06 01:22:05 INFO TaskSetManager: Starting task 0.0 in stage 17.0 (TID 41, quickstart.cloudera, PROCESS_LOCAL, 1290 bytes)
25/11/06 01:22:05 INFO BlockManagerInfo: Added broadcast_22_piece0 in memory on quickstart.cloudera:57065 (size: 2.5 KB, free: 530.2 MB)
25/11/06 01:22:05 INFO BlockManagerInfo: Added broadcast_21_piece0 in memory on quickstart.cloudera:57065 (size: 21.8 KB, free: 530.2 MB)
25/11/06 01:22:05 INFO TaskSetManager: Starting task 1.0 in stage 17.0 (TID 42, quickstart.cloudera, PROCESS_LOCAL, 1290 bytes)
25/11/06 01:22:05 INFO TaskSetManager: Finished task 0.0 in stage 17.0 (TID 41) in 679 ms on quickstart.cloudera (1/2)
25/11/06 01:22:05 INFO DAGScheduler: Stage 17 (map at <console>:21) finished in 0.797 s
25/11/06 01:22:05 INFO DAGScheduler: looking for newly runnable stages
25/11/06 01:22:05 INFO DAGScheduler: running: Set()
25/11/06 01:22:05 INFO DAGScheduler: waiting: Set(Stage 18)
```

```
25/11/06 01:22:05 INFO DAGScheduler: failed: Set()
25/11/06 01:22:05 INFO TaskSetManager: Finished task 1.0 in stage 17.0 (TID 42) in 112 ms on quickstart.cloudera (2/2)
25/11/06 01:22:05 INFO YarnScheduler: Removed TaskSet 17.0, whose tasks have all completed, from pool
25/11/06 01:22:05 INFO DAGScheduler: Missing parents for Stage 18: List()
25/11/06 01:22:05 INFO DAGScheduler: Submitting Stage 18 (ShuffledRDD[28] at aggregateByKey at <console>:23), which is now runnable
25/11/06 01:22:05 INFO MemoryStore: ensureFreeSpace(4840) called with curMem=1546219, maxMem=278302556
25/11/06 01:22:05 INFO MemoryStore: Block broadcast_23 stored as values in memory (estimated size 4.7 KB, free 263.9 MB)
25/11/06 01:22:05 INFO MemoryStore: ensureFreeSpace(2727) called with curMem=1551059, maxMem=278302556
25/11/06 01:22:05 INFO MemoryStore: Block broadcast_23_piece0 stored as bytes in memory (estimated size 2.7 KB, free 263.9 MB)
25/11/06 01:22:05 INFO BlockManagerInfo: Added broadcast_23_piece0 in memory on quickstart.cloudera:54996 (size: 2.7 KB, free: 265.3 MB)
25/11/06 01:22:05 INFO BlockManagerMaster: Updated info of block broadcast_23_piece0
25/11/06 01:22:05 INFO SparkContext: Created broadcast 23 from broadcast at DAGScheduler.scala:839
25/11/06 01:22:05 INFO DAGScheduler: Submitting 2 missing tasks from Stage 18 (ShuffledRDD[28] at aggregateByKey at <console>:23)
25/11/06 01:22:05 INFO YarnScheduler: Adding task set 18.0 with 2 tasks
25/11/06 01:22:05 INFO TaskSetManager: Starting task 0.0 in stage 18.0 (TID 43, quickstart.cloudera, PROCESS_LOCAL, 1056 bytes)
25/11/06 01:22:05 INFO BlockManagerInfo: Added broadcast_23_piece0 in memory on quickstart.cloudera:57065 (size: 2.7 KB, free: 530.2 MB)
25/11/06 01:22:06 INFO MapOutputTrackerMasterActor: Asked to send map output locations for shuffle 7 to sparkExecutor@quickstart.cloudera:40965
25/11/06 01:22:06 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 7 is 160 bytes
25/11/06 01:22:06 INFO TaskSetManager: Starting task 1.0 in stage 18.0 (TID 44, quickstart.cloudera, PROCESS_LOCAL, 1056 bytes)
25/11/06 01:22:06 INFO TaskSetManager: Finished task 0.0 in stage 18.0 (TID 43) in 107 ms on quickstart.cloudera (1/2)
25/11/06 01:22:06 INFO DAGScheduler: Stage 18 (collect at <console>:26) finished in 0.153 s
25/11/06 01:22:06 INFO DAGScheduler: Job 10 finished: collect at <console>:26, took 1.028245 s
res12: Array[(String, Int)] = Array((Shubham,155), (Manas,175))
```

foldByKey:

```
scala> val fold = marks.foldByKey(0)(_ + _)
fold: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[29] at foldByKey at <console>:23

scala> fold.collect()
25/11/06 01:23:23 INFO SparkContext: Starting job: collect at <console>:26
25/11/06 01:23:23 INFO DAGScheduler: Registering RDD 27 (map at <console>:21)
25/11/06 01:23:23 INFO DAGScheduler: Got job 11 (collect at <console>:26) with 2 output partitions (allowLocal=false)
25/11/06 01:23:23 INFO DAGScheduler: Final stage: Stage 20(collect at <console>:26)
25/11/06 01:23:23 INFO DAGScheduler: Parents of final stage: List(Stage 19)
25/11/06 01:23:23 INFO DAGScheduler: Missing parents: List(Stage 19)
25/11/06 01:23:23 INFO DAGScheduler: Submitting Stage 19 (MapPartitionsRDD[27] at map at <console>:21), which has no missing parents
25/11/06 01:23:23 INFO MemoryStore: ensureFreeSpace(4352) called with curMem=1553786, maxMem=278302556
25/11/06 01:23:23 INFO MemoryStore: Block broadcast_24 stored as values in memory (estimated size 4.3 KB, free 263.9 MB)
25/11/06 01:23:23 INFO MemoryStore: ensureFreeSpace(2560) called with curMem=1558138, maxMem=278302556
25/11/06 01:23:23 INFO MemoryStore: Block broadcast_24_piece0 stored as bytes in memory (estimated size 2.5 KB, free 263.9 MB)
25/11/06 01:23:23 INFO BlockManagerInfo: Added broadcast_24_piece0 in memory on quickstart.cloudera:54996 (size: 2.5 KB, free: 265.3 MB)
25/11/06 01:23:23 INFO BlockManagerMaster: Updated info of block broadcast_24_piece0
25/11/06 01:23:23 INFO SparkContext: Created broadcast 24 from broadcast at DAGScheduler.scala:839
25/11/06 01:23:23 INFO DAGScheduler: Submitting 2 missing tasks from Stage 19 (MapPartitionsRDD[27] at map at <console>:21)
25/11/06 01:23:23 INFO YarnScheduler: Adding task set 19.0 with 2 tasks
25/11/06 01:23:23 INFO TaskSetManager: Starting task 0.0 in stage 19.0 (TID 45, quickstart.cloudera, PROCESS_LOCAL, 1290 bytes)
25/11/06 01:23:23 INFO BlockManagerInfo: Added broadcast_24_piece0 in memory on quickstart.cloudera:57065 (size: 2.5 KB, free: 530.2 MB)
25/11/06 01:23:23 INFO TaskSetManager: Starting task 1.0 in stage 19.0 (TID 46, quickstart.cloudera, PROCESS_LOCAL, 1290 bytes)
25/11/06 01:23:23 INFO TaskSetManager: Finished task 0.0 in stage 19.0 (TID 45) in 120 ms on quickstart.cloudera (1/2)
25/11/06 01:23:23 INFO TaskSetManager: Finished task 1.0 in stage 19.0 (TID 46) in 41 ms on quickstart.cloudera (2/2)
25/11/06 01:23:23 INFO YarnScheduler: Removed TaskSet 19.0, whose tasks have all completed, from pool
25/11/06 01:23:23 INFO DAGScheduler: Stage 19 (map at <console>:21) finished in 0.172 s
25/11/06 01:23:23 INFO DAGScheduler: looking for newly runnable stages
25/11/06 01:23:23 INFO DAGScheduler: running: Set()
25/11/06 01:23:23 INFO DAGScheduler: waiting: Set(Stage 20)
25/11/06 01:23:23 INFO DAGScheduler: failed: Set()
25/11/06 01:23:23 INFO DAGScheduler: Missing parents for Stage 20: List()
25/11/06 01:23:23 INFO DAGScheduler: Submitting Stage 20 (ShuffledRDD[29] at foldByKey at <console>:23), which is now runnable
25/11/06 01:23:23 INFO MemoryStore: ensureFreeSpace(4736) called with curMem=1560698, maxMem=278302556
25/11/06 01:23:23 INFO MemoryStore: Block broadcast_25 stored as values in memory (estimated size 4.6 KB, free 263.9 MB)
25/11/06 01:23:23 INFO BlockManager: Removing broadcast 23
25/11/06 01:23:23 INFO BlockManager: Removing block broadcast_23_piece0
25/11/06 01:23:23 INFO MemoryStore: Block broadcast_23_piece0 of size 2727 dropped from memory (free 276739849)
25/11/06 01:23:23 INFO BlockManagerInfo: Removed broadcast_23_piece0 on quickstart.cloudera:54996 in memory (size: 2.7 KB, free: 265.3 MB)
25/11/06 01:23:23 INFO BlockManagerMaster: Updated info of block broadcast_23_piece0
25/11/06 01:23:23 INFO BlockManager: Removing block broadcast_23
25/11/06 01:23:23 INFO MemoryStore: Block broadcast_23 of size 4840 dropped from memory (free 276744689)
25/11/06 01:23:23 INFO BlockManagerInfo: Removed broadcast_23_piece0 on quickstart.cloudera:57065 in memory (size: 2.7 KB, free: 530.2 MB)
25/11/06 01:23:23 INFO ContextCleaner: Cleaned broadcast 23
25/11/06 01:23:23 INFO BlockManager: Removing broadcast 22
25/11/06 01:23:23 INFO BlockManager: Removing block broadcast_22_piece0
25/11/06 01:23:23 INFO MemoryStore: Block broadcast_22_piece0 of size 2570 dropped from memory (free 276747259)
25/11/06 01:23:23 INFO BlockManagerInfo: Removed broadcast_22_piece0 on quickstart.cloudera:54996 in memory (size: 2.5 KB, free: 265.3 MB)
25/11/06 01:23:23 INFO BlockManagerMaster: Updated info of block broadcast_22_piece0
25/11/06 01:23:23 INFO BlockManager: Removing block broadcast_22
25/11/06 01:23:23 INFO MemoryStore: Block broadcast_22 of size 4464 dropped from memory (free 276751723)
25/11/06 01:23:23 INFO BlockManagerInfo: Removed broadcast_22_piece0 on quickstart.cloudera:57065 in memory (size: 2.5 KB, free: 530.2 MB)
25/11/06 01:23:23 INFO MemoryStore: ensureFreeSpace(2709) called with curMem=1550833, maxMem=278302556
25/11/06 01:23:23 INFO MemoryStore: Block broadcast_25_piece0 stored as bytes in memory (estimated size 2.6 KB, free 263.9 MB)
25/11/06 01:23:23 INFO BlockManagerInfo: Added broadcast_25_piece0 in memory on quickstart.cloudera:54996 (size: 2.6 KB, free: 265.3 MB)
25/11/06 01:23:23 INFO ContextCleaner: Cleaned broadcast 22
25/11/06 01:23:23 INFO BlockManagerMaster: Updated info of block broadcast_25_piece0
25/11/06 01:23:23 INFO SparkContext: Created broadcast 25 from broadcast at DAGScheduler.scala:839
25/11/06 01:23:23 INFO DAGScheduler: Submitting 2 missing tasks from Stage 20 (ShuffledRDD[29] at foldByKey at <console>:23)
25/11/06 01:23:23 INFO YarnScheduler: Adding task set 20.0 with 2 tasks
25/11/06 01:23:23 INFO TaskSetManager: Starting task 0.0 in stage 20.0 (TID 47, quickstart.cloudera, PROCESS_LOCAL, 1056 bytes)
25/11/06 01:23:23 INFO BlockManagerInfo: Added broadcast_25_piece0 in memory on quickstart.cloudera:57065 (size: 2.6 KB, free: 530.2 MB)
25/11/06 01:23:23 INFO MapOutputTrackerMasterActor: Asked to send map output locations for shuffle 8 to sparkExecutor@quickstart.cloudera:40965
25/11/06 01:23:23 INFO MapOutputTrackerMaster: Size of output statuses for shuffle 8 is 160 bytes
25/11/06 01:23:23 INFO TaskSetManager: Starting task 1.0 in stage 20.0 (TID 48, quickstart.cloudera, PROCESS_LOCAL, 1056 bytes)
25/11/06 01:23:23 INFO TaskSetManager: Finished task 0.0 in stage 20.0 (TID 47) in 75 ms on quickstart.cloudera (1/2)
25/11/06 01:23:23 INFO DAGScheduler: Stage 20 (collect at <console>:26) finished in 0.131 s
25/11/06 01:23:23 INFO DAGScheduler: Job 11 finished: collect at <console>:26, took 0.391970 s
res13: Array[(String, Int)] = Array((Shubham,155), (Manas,175))
```

Read from Local File:

BIGDATA

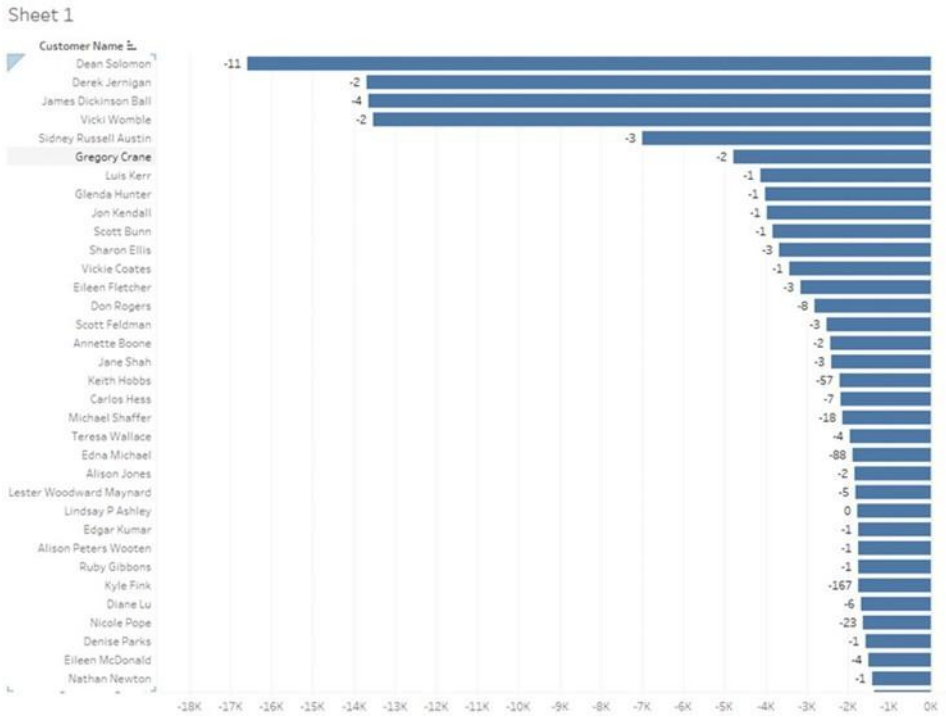
```
scala> localRdd.collect().foreach(println)
25/11/06 01:24:40 INFO FileInputFormat: Total input paths to process : 1
25/11/06 01:24:40 INFO SparkContext: Starting job: collect at <console>:24
25/11/06 01:24:40 INFO DAGScheduler: Got job 12 (collect at <console>:24) with 2 output partitions (allowLocal=false)
25/11/06 01:24:40 INFO DAGScheduler: Final stage: Stage 21(collect at <console>:24)
25/11/06 01:24:40 INFO DAGScheduler: Parents of final stage: List()
25/11/06 01:24:40 INFO DAGScheduler: Missing parents: List()
25/11/06 01:24:40 INFO DAGScheduler: Submitting Stage 21 (file:/home/cloudera/file.txt MapPartitionsRDD[31] at textFile at <console>:21), which has no missing parent
s
25/11/06 01:24:40 INFO MemoryStore: ensureFreeSpace(2664) called with curMem=1858930, maxMem=278302556
25/11/06 01:24:40 INFO MemoryStore: Block broadcast_27 stored as values in memory (estimated size 2.6 KB, free 263.6 MB)
25/11/06 01:24:40 INFO MemoryStore: ensureFreeSpace(1636) called with curMem=1861594, maxMem=278302556
25/11/06 01:24:40 INFO MemoryStore: Block broadcast_27_piece0 stored as bytes in memory (estimated size 1636.0 B, free 263.6 MB)
25/11/06 01:24:40 INFO BlockManagerInfo: Added broadcast_27_piece0 in memory on quickstart.cloudera:54996 (size: 1636.0 B, free: 265.3 MB)
25/11/06 01:24:40 INFO BlockManagerMaster: Updated info of block broadcast_27_piece0
25/11/06 01:24:40 INFO SparkContext: Created broadcast 27 from broadcast at DAGScheduler.scala:839
25/11/06 01:24:40 INFO DAGScheduler: Submitting 2 missing tasks from Stage 21 (file:/home/cloudera/file.txt MapPartitionsRDD[31] at textFile at <console>:21)
25/11/06 01:24:40 INFO YarnScheduler: Adding task set 21.0 with 2 tasks
25/11/06 01:24:40 INFO TaskSetManager: Starting task 0.0 in stage 21.0 (TID 49, quickstart.cloudera, PROCESS_LOCAL, 1292 bytes)
25/11/06 01:24:40 INFO BlockManagerInfo: Added broadcast_27_piece0 in memory on quickstart.cloudera:57065 (size: 1636.0 B, free: 530.2 MB)
25/11/06 01:24:40 INFO BlockManagerInfo: Added broadcast_26_piece0 in memory on quickstart.cloudera:57065 (size: 21.8 KB, free: 530.1 MB)
25/11/06 01:24:40 INFO TaskSetManager: Starting task 1.0 in stage 21.0 (TID 50, quickstart.cloudera, PROCESS_LOCAL, 1292 bytes)
25/11/06 01:24:40 INFO TaskSetManager: Finished task 0.0 in stage 21.0 (TID 49) in 219 ms on quickstart.cloudera (1/2)
25/11/06 01:24:40 INFO DAGScheduler: Stage 21 (collect at <console>:24) finished in 0.277 s
25/11/06 01:24:40 INFO DAGScheduler: Job 12 finished: collect at <console>:24, took 0.327347 s
25/11/06 01:24:40 INFO TaskSetManager: Finished task 1.0 in stage 21.0 (TID 50) in 58 ms on quickstart.cloudera (2/2)
25/11/06 01:24:40 INFO YarnScheduler: Removed TaskSet 21.0, whose tasks have all completed, from pool
hello world
hello spark
spark is fast
```

7. Visualization using Tableau: Tableau: Tool Overview, Importing Data, Analyzing with Charts, Creating Dashboards, working with maps.

Analysis Operations

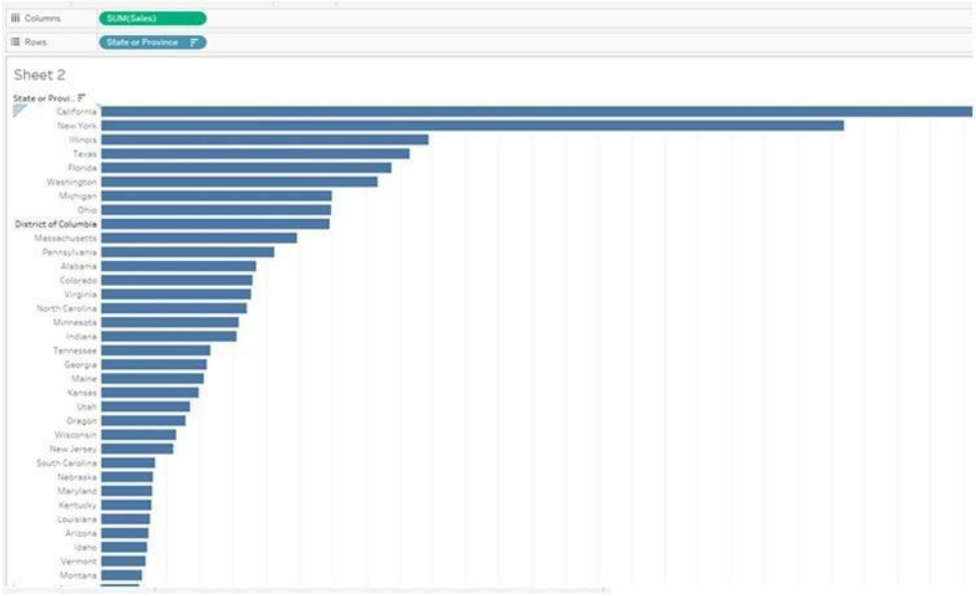
Q1. Find the customer with the highest overall profit. What is his/her profit ratio?

- Import superstoreus2015.xlsx dataset.
- Create Calculated Field:
- Profit Ratio = (SUM([Profit]) / SUM([Sales]))
- Drag Customer Name to Rows, Profit to Columns. Sort by Profit.
- Show Profit Ratio in Marks label



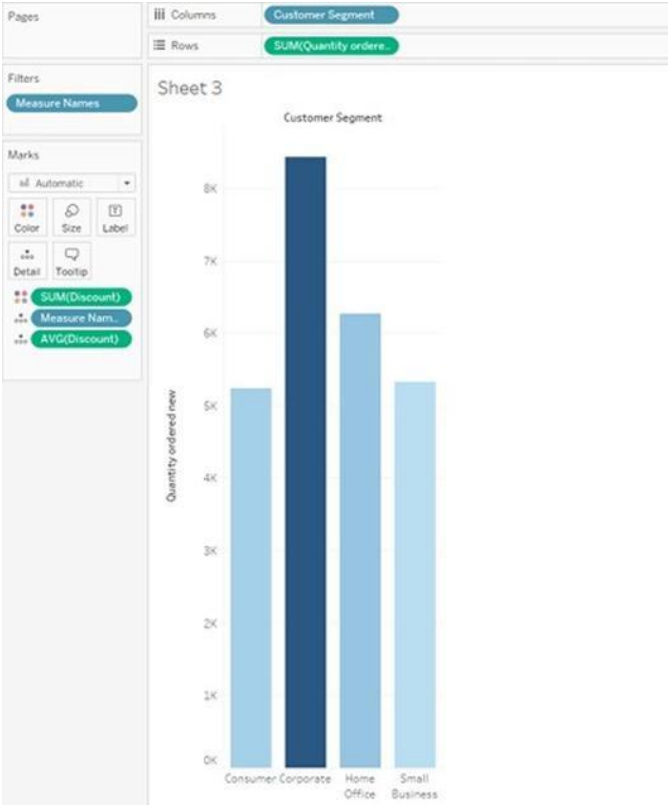
Q2. Which state has the highest Sales (Sum)? What is the total Sales for that state?

- Drag State to Rows, Sales to Columns.
- Sort descending



Q3. Which customer segment has both the highest order quantity and average discount rate?

- Drag Segment to Rows.
- Add Order Quantity and Discount to Measures.
- Analyze visually.



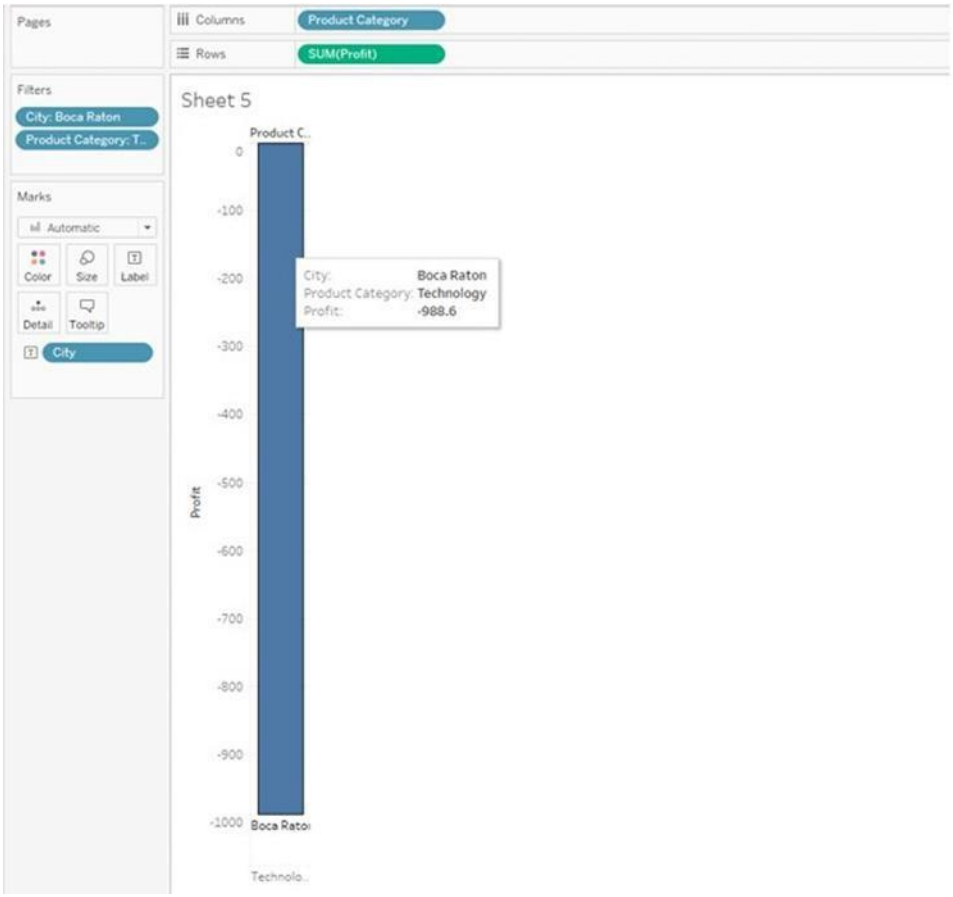
Q4. Which Product Category has the highest total Sales? Which Product Category has the worst Profit?

- Create a Bar Chart with Category vs Sales.
- Add Profit as color scale

Q5. What was the Profit on Technology in Boca Raton (City)?

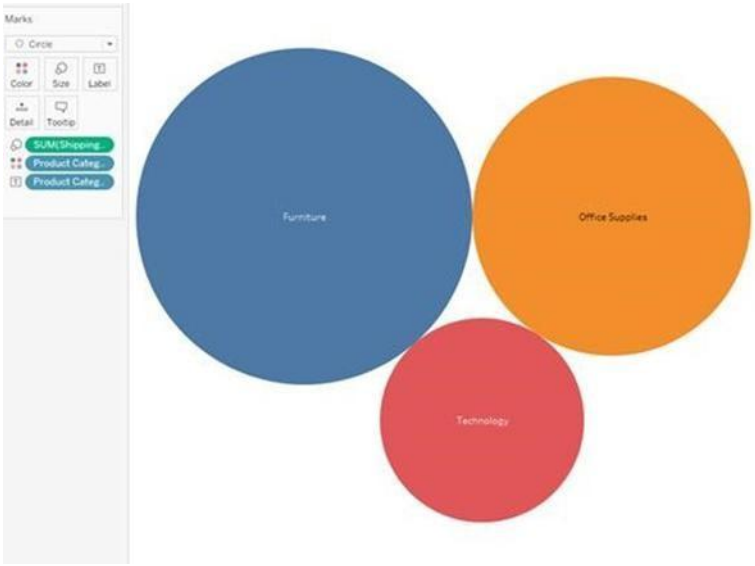
- Add City filter.
- Filter for “Boca Raton”.
- Show Profit.

BIGDATA



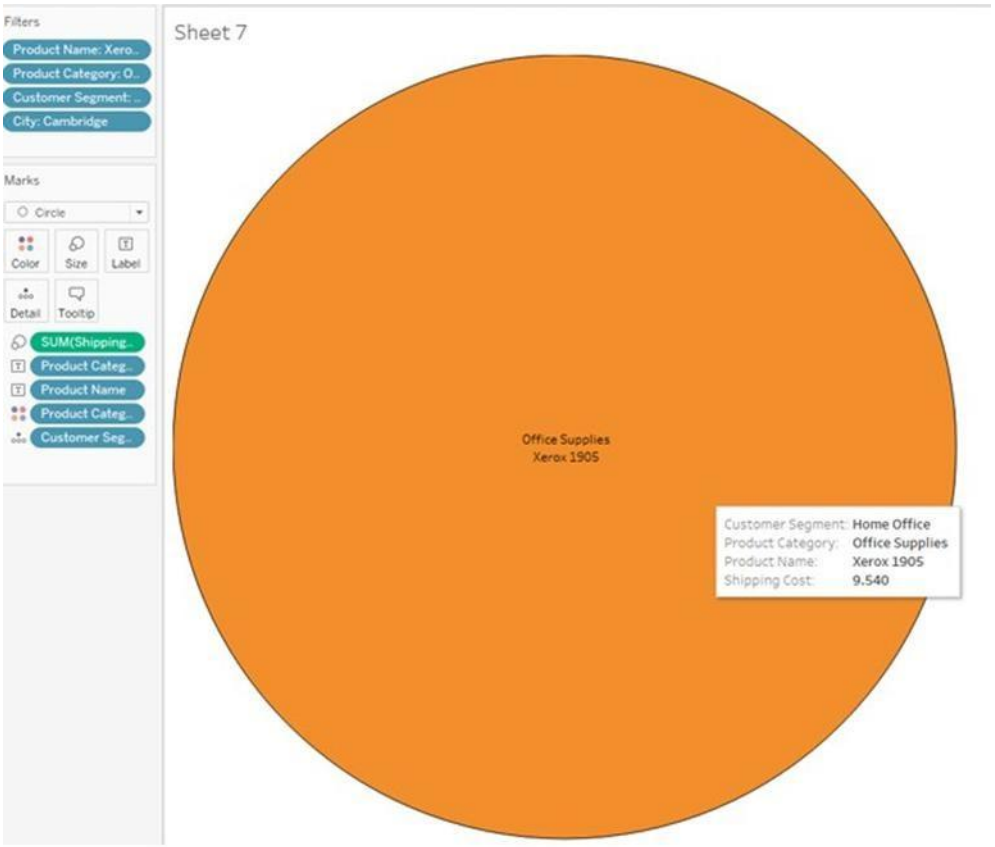
Q6. Which Product Department has the highest Shipping Costs?

- Create Packed Bubble Chart with Department.
- Use Shipping Cost as bubble size.



Q7. What was the shipping cost of Office Supplies for Xerox 1905 in the Home Customer Segment in Cambridge?

- Apply filters: Product → Xerox 1905, Category → Office Supplies, Segment → Home, City → Cambridge.

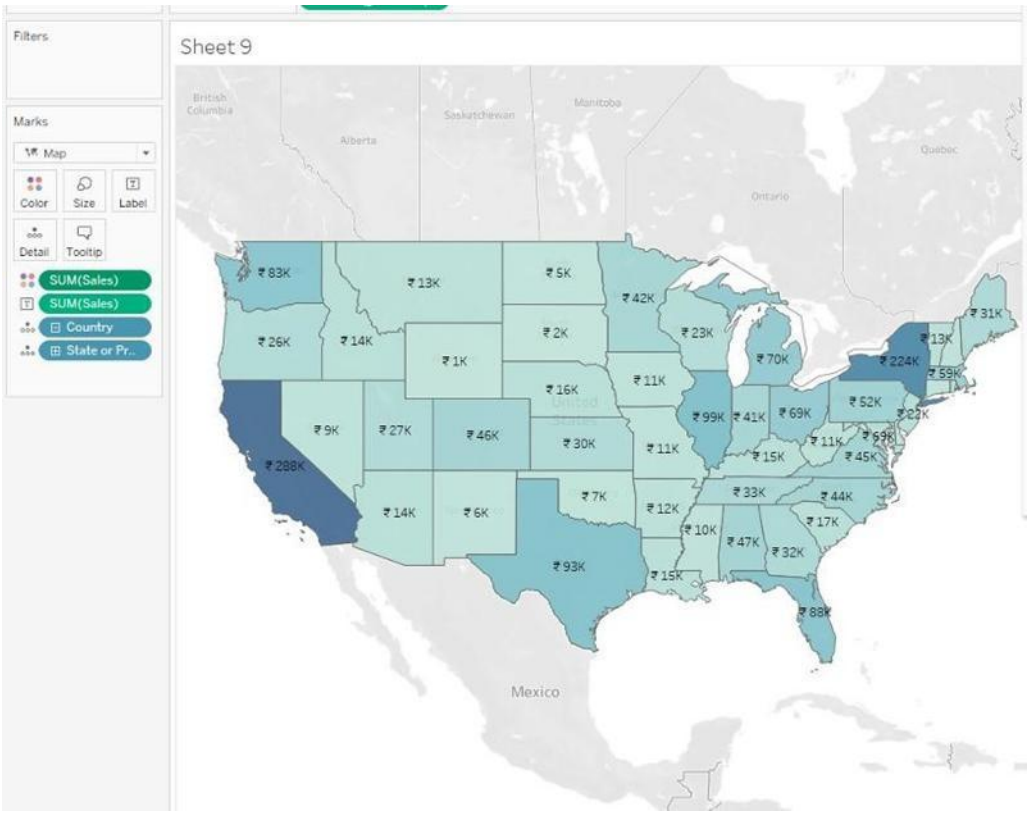


Preparing Maps

Dataset: Superstore

1. Create Geographic Hierarchy (Country → State → City → Postal Code)
   A. In the Data Pane, right-click on Country → select Hierarchy → Create Hierarchy.
   B. Name it (e.g., Geography). 3. Drag State, City, and Postal Code into the hierarchy (below Country in order).
   C. Now you can drill down from Country → State → City → Postal Code.

2. Build a Basic Map of Sales by State
   A. Double-click on State → Tableau will create a map.
   B. From Measures, drag Sales → drop it on Color (Marks card). Now each state is colored by its Sales
   C. Drag Sales again → drop it on Label (so values appear on the map).
   D. Right-click on Sales → Default Properties → Number Format → Currency / Custom (No decimals, in K).



3. Add Profit Ratio as Tooltip

A. Create a calculated field.

B. Profit Ratio = SUM([Profit]) / SUM([Sales])

C. Drag this field onto the Tooltip shelf (Marks card).

D. Now when you hover over a state, you will see both Sales and Profit Ratio.

E. Show Profit Ratio of each state as tooltip on map

F. Show Profit ratio for Grip Envelop products

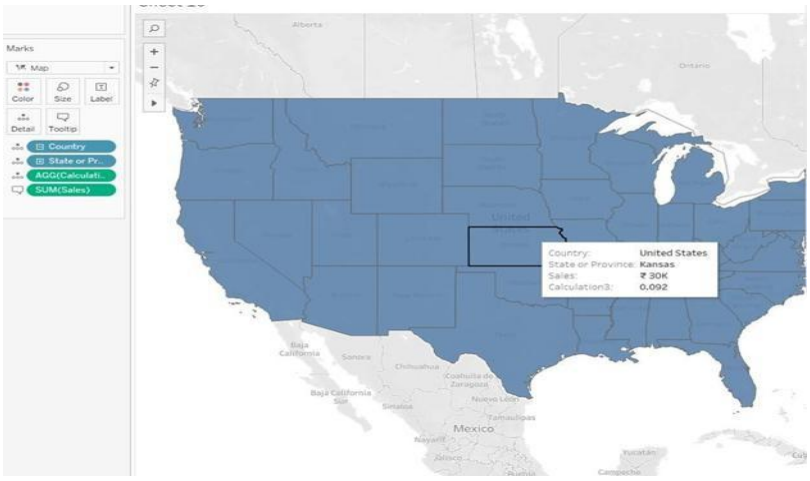
G. Identify Unprofitable States in Technology Surrounded by Profitable Ones (e.g., Nevada)

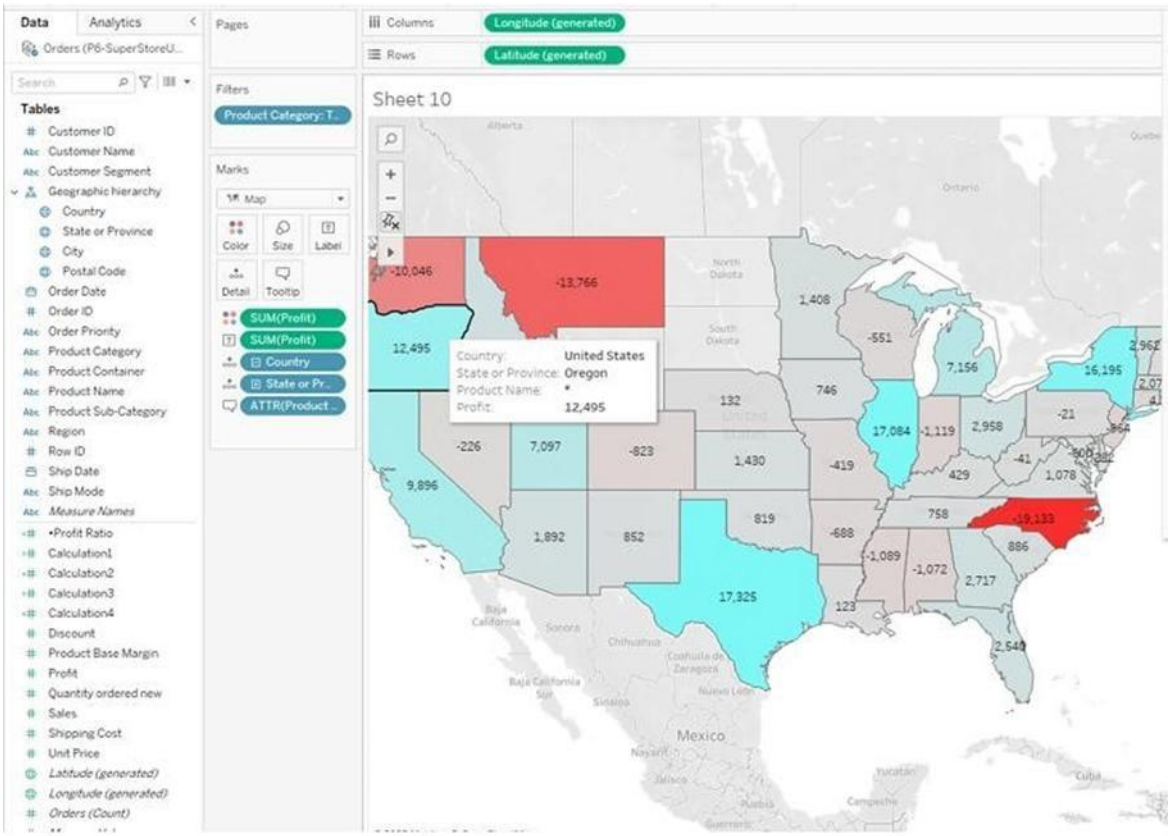
a) 1. Drag Category to Filters → select only Technology.

b) Drag Category to Filters → select only Technology.

c) Drag Profit to Color (Marks card). o Positive profit → blue/green, Negative profit → red.

d) Add Profit also to Label if you want exact values. 4. Look at the map → Nevada will appear red (loss) but surrounded by green states (profit).





4. Preparing Reports

Dataset: Superstore

- Report: Product category-wise sales.
- Report: Region-wise product sales.
- Report: State-wise sales.
- Example Question: What is % of total Sales for Home Office segment in 2015?
- Example Question:
  a) Find Top 10 Product Names by Sales in each region.
  b) Which product is ranked #2 in Central & West in 2015?

Conclusion: Successfully performed Visualization in Tableau with charts, dashboards, stories, maps, and reports.