B.M.S. COLLEGE OF ENGINEERING BENGALURU
Autonomous Institute, Affiliated to VTU

Lab Record

Big Data Analytics

Submitted in partial fulfillment for the 6th Semester Laboratory

Bachelor of Technology
in
Computer Science and Engineering

Submitted by:

Ankit Kesar

1BM18CS150

Department of Computer Science and Engineering
B.M.S. College of Engineering
Bull Temple Road, Basavanagudi, Bangalore 560 019
Mar-June 2021

B.M.S. COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Big Data Analytics (20CS6PEBDA) laboratory has been carried out by Ankit Kesar (1BM18CS150) during the 6th Semester Mar-June-2021.

Signature of the Faculty Incharge:

Bhoomika :

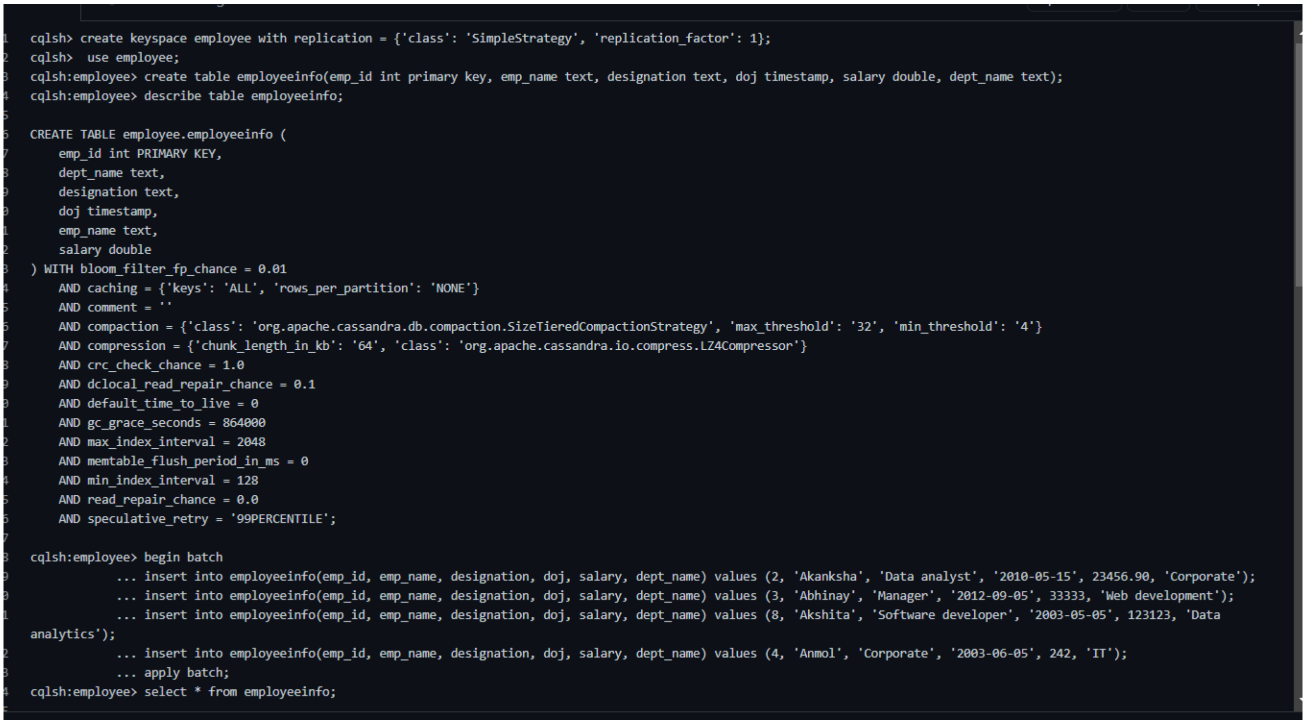
Department of Computer Science and Engineering
B.M.S. College of Engineering, Bangalore

Table of Contents

Program 1. Perform the following DB operations using Cassandra.

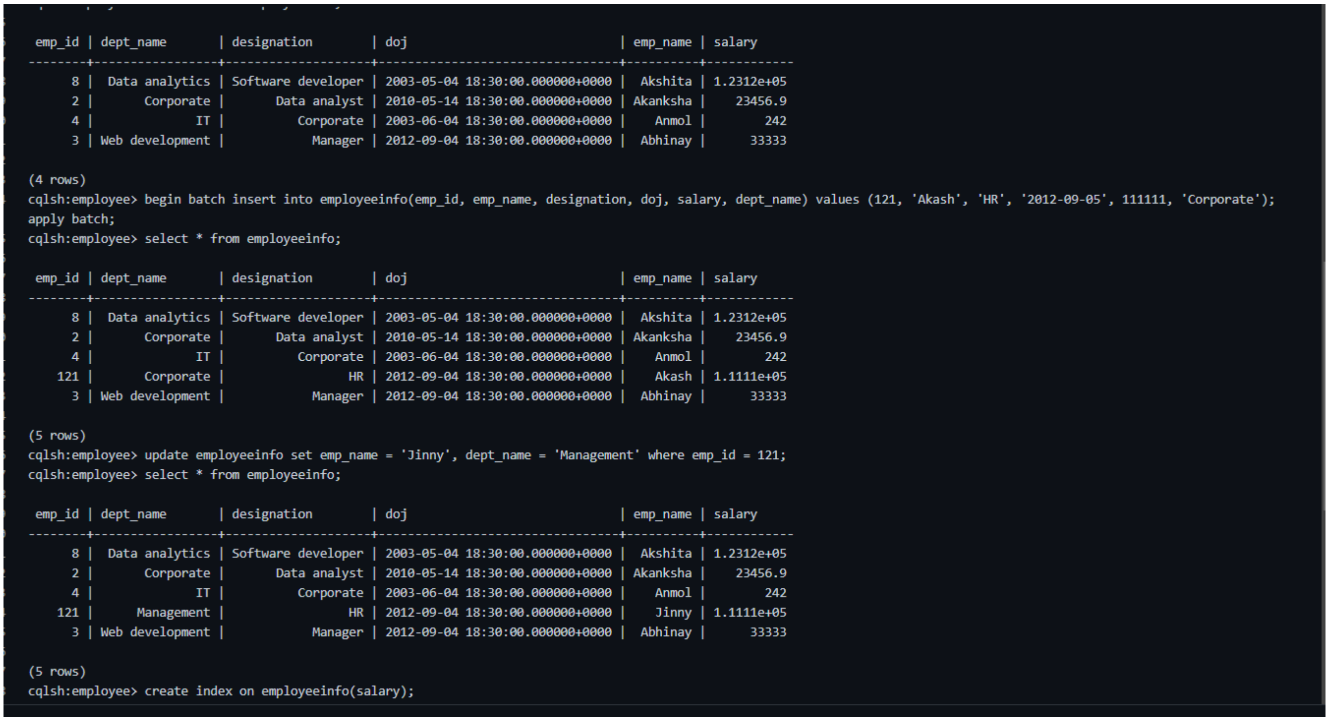
1. Create a keyspace by name Employee
2. Create a column family by name Employee-Info with attributes Emp_Id Primary Key, Emp_Name, Designation, Date_of_Joining, Salary, Dept_Name
3. Insert the values into the table in batch
4. Update Employee name and Department of Emp-Id 121
5. Sort the details of Employee records based on salary
6. Alter the schema of the table Employee_Info to add a column Projects which stores a set of Projects done by the corresponding Employee.
7. Update the altered table to add project names.

8 Create a TTL of 15 seconds to display the values of Employees.



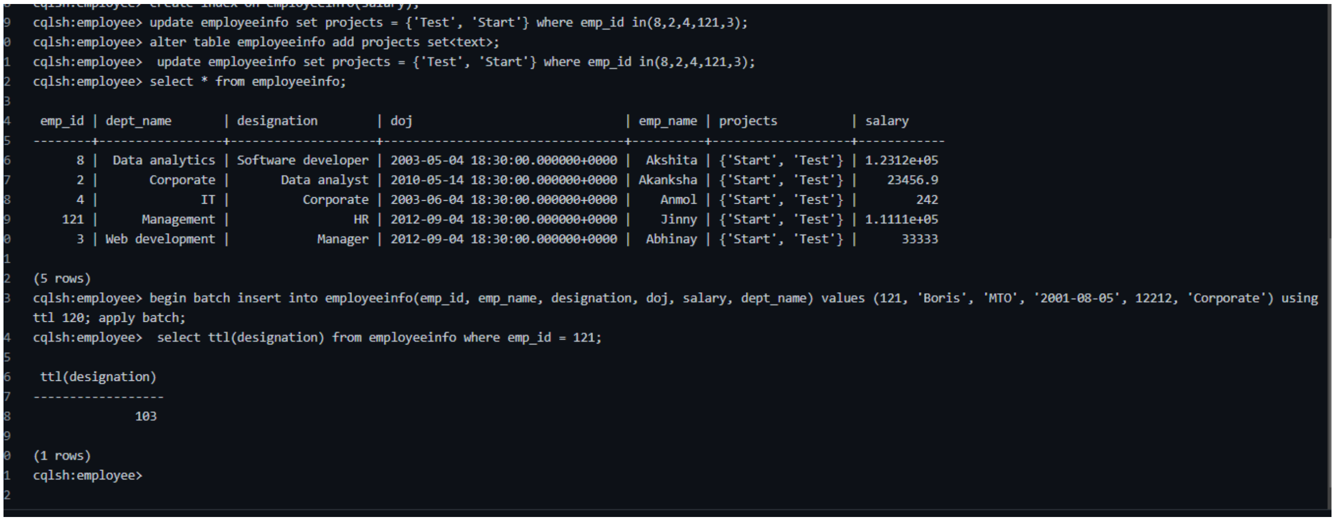
```
cqlsh> create keyspace employee with replication = {'class': 'SimpleStrategy', 'replication_factor': 1};
cqlsh>  use employee;
cqlsh:employee> create table employeeinfo(emp_id int primary key, emp_name text, designation text, doj timestamp, salary double, dept_name text);
cqlsh:employee> describe table employeeinfo;

CREATE TABLE employee.employeeinfo (
    emp_id int PRIMARY KEY,
    dept_name text,
    designation text,
    doj timestamp,
    emp_name text,
    salary double
) WITH bloom_filter_fp_chance = 0.01
    AND caching = {'keys': 'ALL', 'rows_per_partition': 'NONE'}
    AND comment = ''
    AND compaction = {'class': 'org.apache.cassandra.db.compaction.SizeTieredCompactionStrategy', 'max_threshold': '32', 'min_threshold': '4'}
    AND compression = {'chunk_length_in_kb': '64', 'class': 'org.apache.cassandra.io.compress.LZ4Compressor'}
    AND crc_check_chance = 1.0
    AND dclocal_read_repair_chance = 0.1
    AND default_time_to_live = 0
    AND gc_grace_seconds = 864000
    AND max_index_interval = 2048
    AND memtable_flush_period_in_ms = 0
    AND min_index_interval = 128
    AND read_repair_chance = 0.0
    AND speculative_retry = '99PERCENTILE';

cqlsh:employee> begin batch
            ... insert into employeeinfo(emp_id, emp_name, designation, doj, salary, dept_name) values (2, 'Akanksha', 'Data analyst', '2010-05-15', 23456.90, 'Corporate');
            ... insert into employeeinfo(emp_id, emp_name, designation, doj, salary, dept_name) values (3, 'Abhinay', 'Manager', '2012-09-05', 33333, 'Web development');
            ... insert into employeeinfo(emp_id, emp_name, designation, doj, salary, dept_name) values (8, 'Akshita', 'Software developer', '2003-05-05', 123123, 'Data
analytics');
            ... insert into employeeinfo(emp_id, emp_name, designation, doj, salary, dept_name) values (4, 'Anmol', 'Corporate', '2003-06-05', 242, 'IT');
            ... apply batch;
cqlsh:employee> select * from employeeinfo;
```

Fig 1.1



```
 emp_id | dept_name       | designation        | doj                             | emp_name | salary
--------+-----------------+--------------------+---------------------------------+----------+------------
      8 |  Data analytics | Software developer | 2003-05-04 18:30:00.000000+0000 |  Akshita | 1.2312e+05
      2 |       Corporate |       Data analyst | 2010-05-14 18:30:00.000000+0000 | Akanksha |    23456.9
      4 |              IT |          Corporate | 2003-06-04 18:30:00.000000+0000 |    Anmol |        242
      3 | Web development |            Manager | 2012-09-04 18:30:00.000000+0000 |  Abhinay |      33333

(4 rows)
cqlsh:employee> begin batch insert into employeeinfo(emp_id, emp_name, designation, doj, salary, dept_name) values (121, 'Akash', 'HR', '2012-09-05', 111111, 'Corporate');
apply batch;
cqlsh:employee> select * from employeeinfo;

 emp_id | dept_name       | designation        | doj                             | emp_name | salary
--------+-----------------+--------------------+---------------------------------+----------+------------
      8 |  Data analytics | Software developer | 2003-05-04 18:30:00.000000+0000 |  Akshita | 1.2312e+05
      2 |       Corporate |       Data analyst | 2010-05-14 18:30:00.000000+0000 | Akanksha |    23456.9
      4 |              IT |          Corporate | 2003-06-04 18:30:00.000000+0000 |    Anmol |        242
    121 |       Corporate |                 HR | 2012-09-04 18:30:00.000000+0000 |    Akash | 1.1111e+05
      3 | Web development |            Manager | 2012-09-04 18:30:00.000000+0000 |  Abhinay |      33333

(5 rows)
cqlsh:employee> update employeeinfo set emp_name = 'Jinny', dept_name = 'Management' where emp_id = 121;
cqlsh:employee> select * from employeeinfo;

 emp_id | dept_name       | designation        | doj                             | emp_name | salary
--------+-----------------+--------------------+---------------------------------+----------+------------
      8 |  Data analytics | Software developer | 2003-05-04 18:30:00.000000+0000 |  Akshita | 1.2312e+05
      2 |       Corporate |       Data analyst | 2010-05-14 18:30:00.000000+0000 | Akanksha |    23456.9
      4 |              IT |          Corporate | 2003-06-04 18:30:00.000000+0000 |    Anmol |        242
    121 |      Management |                 HR | 2012-09-04 18:30:00.000000+0000 |    Jinny | 1.1111e+05
      3 | Web development |            Manager | 2012-09-04 18:30:00.000000+0000 |  Abhinay |      33333

(5 rows)
cqlsh:employee> create index on employeeinfo(salary);
```

Fig 1.2

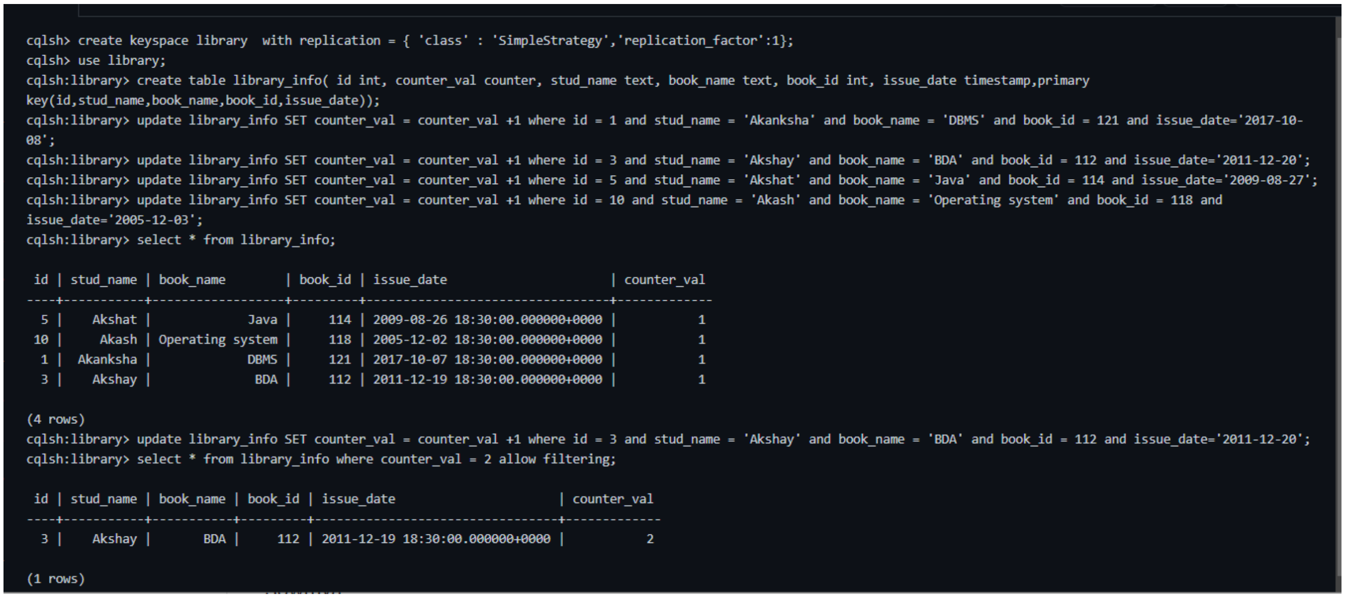


```
cqlsh:employee> create index on employeeinfo(salary);
cqlsh:employee> update employeeinfo set projects = {'Test', 'Start'} where emp_id in(8,2,4,121,3);
cqlsh:employee> alter table employeeinfo add projects set<text>;
cqlsh:employee>  update employeeinfo set projects = {'Test', 'Start'} where emp_id in(8,2,4,121,3);
cqlsh:employee> select * from employeeinfo;

 emp_id | dept_name       | designation        | doj                             | emp_name | projects           | salary
--------+-----------------+--------------------+---------------------------------+----------+--------------------+------------
      8 |  Data analytics | Software developer | 2003-05-04 18:30:00.000000+0000 |  Akshita | {'Start', 'Test'} | 1.2312e+05
      2 |       Corporate |       Data analyst | 2010-05-14 18:30:00.000000+0000 | Akanksha | {'Start', 'Test'} |    23456.9
      4 |              IT |          Corporate | 2003-06-04 18:30:00.000000+0000 |    Anmol | {'Start', 'Test'} |        242
    121 |      Management |                 HR | 2012-09-04 18:30:00.000000+0000 |    Jinny | {'Start', 'Test'} | 1.1111e+05
      3 | Web development |            Manager | 2012-09-04 18:30:00.000000+0000 |  Abhinay | {'Start', 'Test'} |      33333

(5 rows)
cqlsh:employee> begin batch insert into employeeinfo(emp_id, emp_name, designation, doj, salary, dept_name) values (121, 'Boris', 'MTO', '2001-08-05', 12212, 'Corporate') using
ttl 120; apply batch;
cqlsh:employee>  select ttl(designation) from employeeinfo where emp_id = 121;

 ttl(designation)
------------------
              103

(1 rows)
cqlsh:employee>
```

Fig 1.3

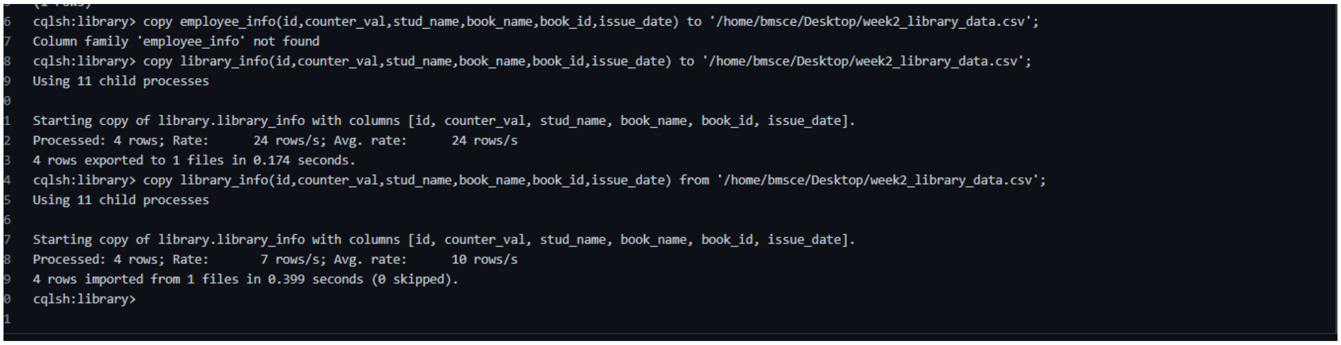
Program 2. Perform the following DB operations using Cassandra.

1 Create a keyspace by name Library
2. Create a column family by name Library-Info with attributes
Stud_Id Primary Key,
Counter_value of type Counter,
Stud_Name, Book-Name, Book-Id, Date_of_issue
3. Insert the values into the table in batch
4. Display the details of the table created and increase the value of the counter
5. Write a query to show that a student with id 112 has taken a book "BDA" 2 times.
6. Export the created column to a csv file
7. Import a given csv dataset from local file system into Cassandra column family



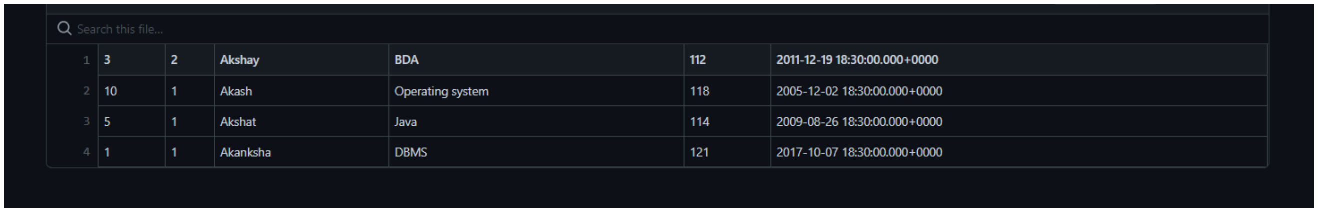
```
cqlsh> create keyspace library  with replication = { 'class' : 'SimpleStrategy','replication_factor':1};
cqlsh> use library;
cqlsh:library> create table library_info( id int, counter_val counter, stud_name text, book_name text, book_id int, issue_date timestamp,primary
key(id,stud_name,book_name,book_id,issue_date));
cqlsh:library> update library_info SET counter_val = counter_val +1 where id = 1 and stud_name = 'Akanksha' and book_name = 'DBMS' and book_id = 121 and issue_date='2017-10-
08';
cqlsh:library> update library_info SET counter_val = counter_val +1 where id = 3 and stud_name = 'Akshay' and book_name = 'BDA' and book_id = 112 and issue_date='2011-12-20';
cqlsh:library> update library_info SET counter_val = counter_val +1 where id = 5 and stud_name = 'Akshat' and book_name = 'Java' and book_id = 114 and issue_date='2009-08-27';
cqlsh:library> update library_info SET counter_val = counter_val +1 where id = 10 and stud_name = 'Akash' and book_name = 'Operating system' and book_id = 118 and
issue_date='2005-12-03';
cqlsh:library> select * from library_info;

 id | stud_name | book_name        | book_id | issue_date                      | counter_val
----+-----------+------------------+---------+---------------------------------+-------------
  5 |    Akshat |             Java |     114 | 2009-08-26 18:30:00.000000+0000 |           1
 10 |     Akash | Operating system |     118 | 2005-12-02 18:30:00.000000+0000 |           1
  1 |  Akanksha |             DBMS |     121 | 2017-10-07 18:30:00.000000+0000 |           1
  3 |    Akshay |              BDA |     112 | 2011-12-19 18:30:00.000000+0000 |           1

(4 rows)
cqlsh:library> update library_info SET counter_val = counter_val +1 where id = 3 and stud_name = 'Akshay' and book_name = 'BDA' and book_id = 112 and issue_date='2011-12-20';
cqlsh:library> select * from library_info where counter_val = 2 allow filtering;

 id | stud_name | book_name | book_id | issue_date                      | counter_val
----+-----------+-----------+---------+---------------------------------+-------------
  3 |    Akshay |       BDA |     112 | 2011-12-19 18:30:00.000000+0000 |           2

(1 rows)
```

Fig 2.1



```
cqlsh:library> copy employee_info(id,counter_val,stud_name,book_name,book_id,issue_date) to '/home/bmsce/Desktop/week2_library_data.csv';
Column family 'employee_info' not found
cqlsh:library> copy library_info(id,counter_val,stud_name,book_name,book_id,issue_date) to '/home/bmsce/Desktop/week2_library_data.csv';
Using 11 child processes

Starting copy of library.library_info with columns [id, counter_val, stud_name, book_name, book_id, issue_date].
Processed: 4 rows; Rate:      24 rows/s; Avg. rate:      24 rows/s
4 rows exported to 1 files in 0.174 seconds.
cqlsh:library> copy library_info(id,counter_val,stud_name,book_name,book_id,issue_date) from '/home/bmsce/Desktop/week2_library_data.csv';
Using 11 child processes

Starting copy of library.library_info with columns [id, counter_val, stud_name, book_name, book_id, issue_date].
Processed: 4 rows; Rate:       7 rows/s; Avg. rate:      10 rows/s
4 rows imported from 1 files in 0.399 seconds (0 skipped).
cqlsh:library>
```

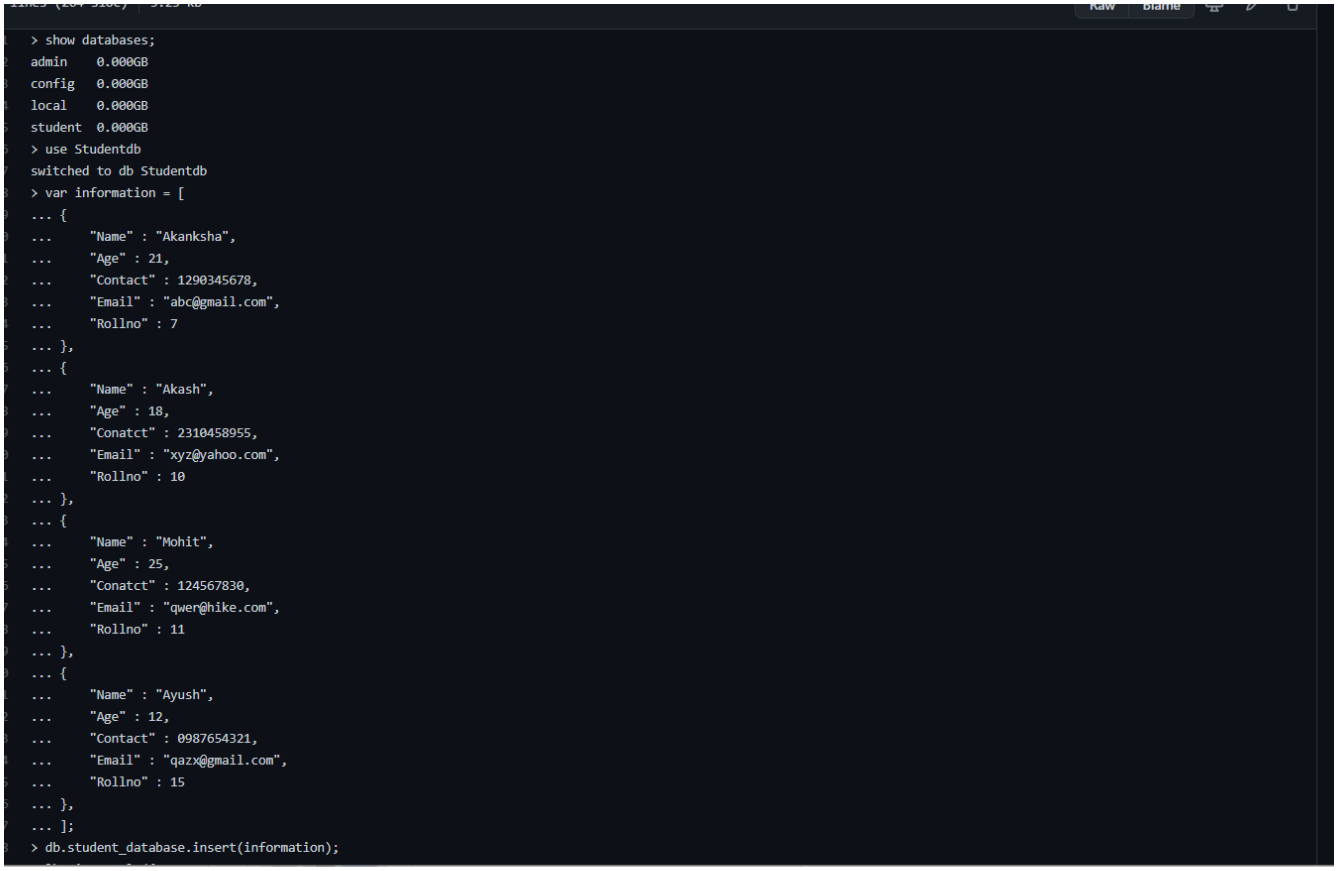
Fig 2.2



3	2	Akshay	BDA	112	2011-12-19 18:30:00.000+0000
10	1	Akash	Operating system	118	2005-12-02 18:30:00.000+0000
5	1	Akshat	Java	114	2009-08-26 18:30:00.000+0000
1	1	Akanksha	DBMS	121	2017-10-07 18:30:00.000+0000

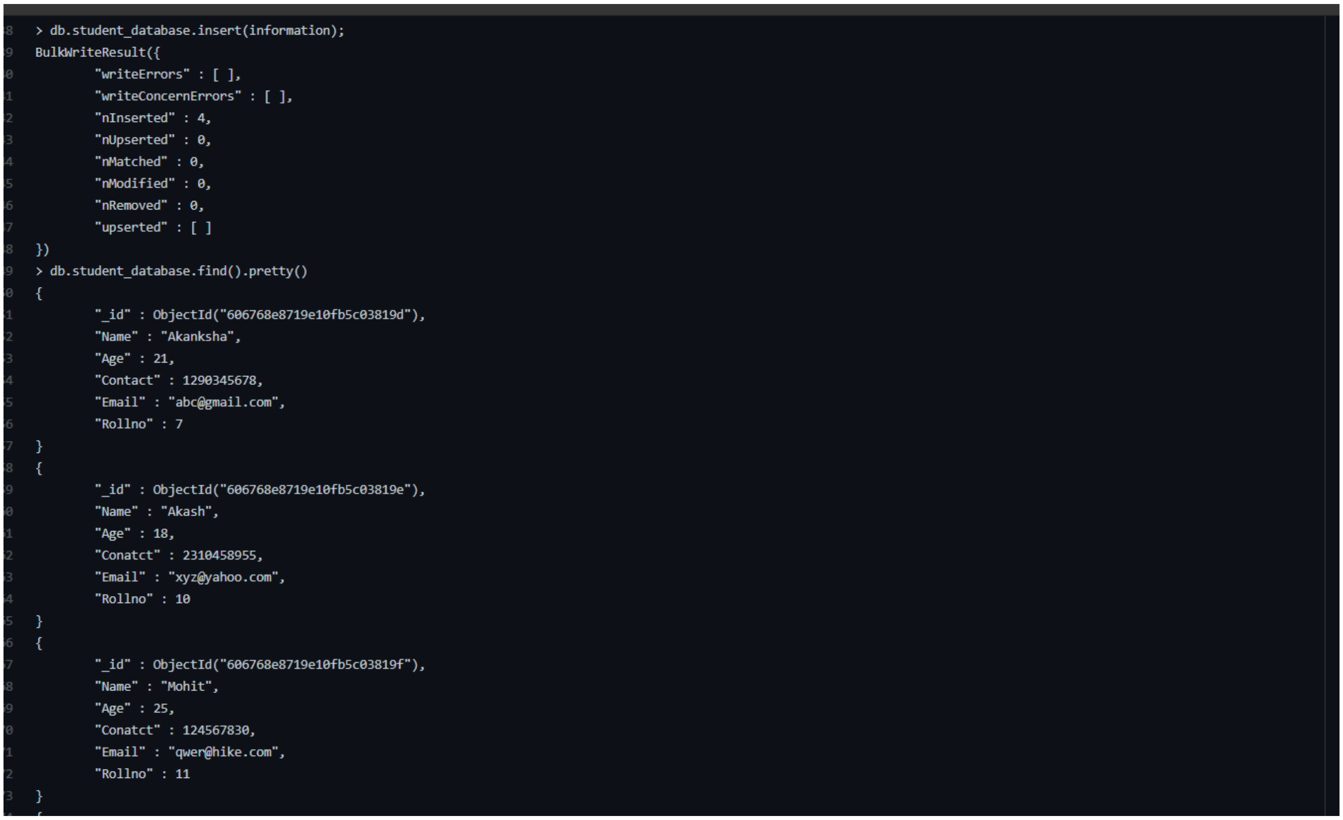
Fig 2.3

Program 3. Perform the following DB operations using MongoDB.1. Create a database “Student” with the following attributes Rollno, Age, ContactNo, Email-Id.2. Insert appropriate values3. Write query to update Email-Id of a student with rollno 10.4. Replace the student name from “ABC” to “FEM” of rollno 11.5. Export the created table into local file system6. Drop the table7. Import a given csv dataset from local file system into mongodb collection.



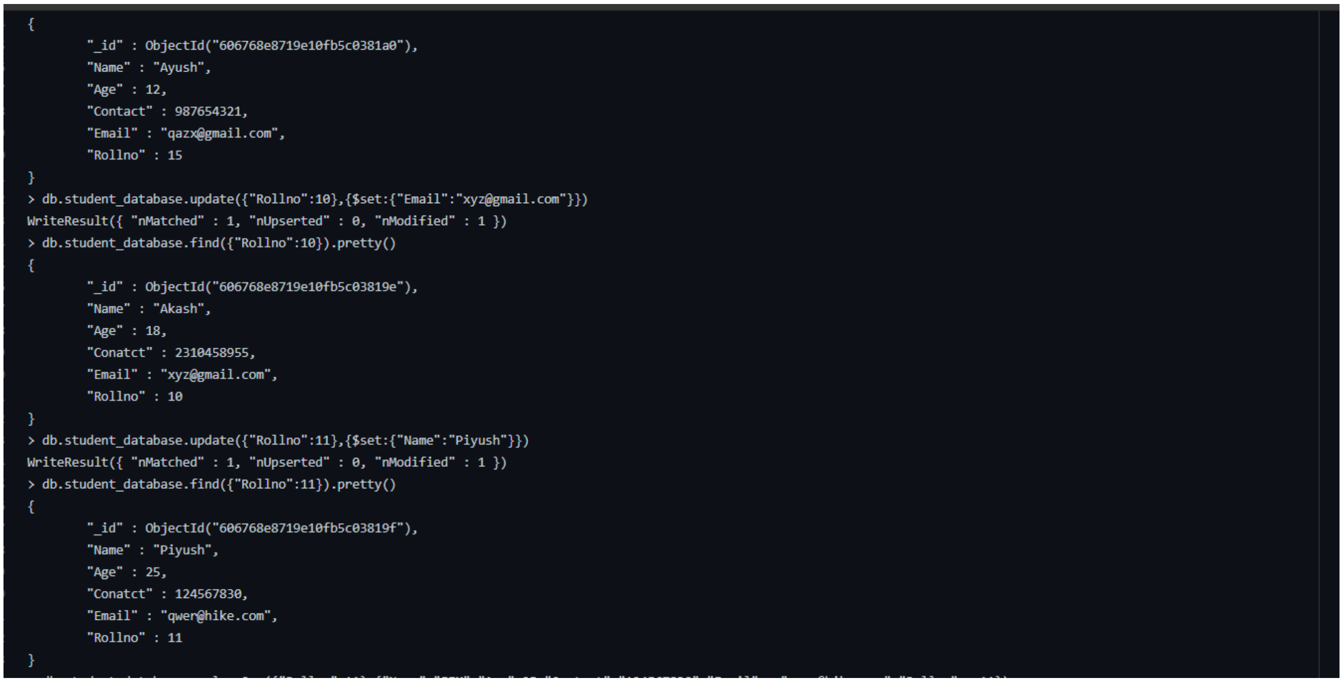
```
> show databases;
admin    0.000GB
config   0.000GB
local    0.000GB
student  0.000GB
> use Studentdb
switched to db Studentdb
> var information = [
... {
...     "Name" : "Akanksha",
...     "Age" : 21,
...     "Contact" : 1290345678,
...     "Email" : "abc@gmail.com",
...     "Rollno" : 7
... },
... {
...     "Name" : "Akash",
...     "Age" : 18,
...     "Conatct" : 2310458955,
...     "Email" : "xyz@yahoo.com",
...     "Rollno" : 10
... },
... {
...     "Name" : "Mohit",
...     "Age" : 25,
...     "Conatct" : 124567830,
...     "Email" : "qwer@hike.com",
...     "Rollno" : 11
... },
... {
...     "Name" : "Ayush",
...     "Age" : 12,
...     "Contact" : 0987654321,
...     "Email" : "qazx@gmail.com",
...     "Rollno" : 15
... },
... ];
> db.student_database.insert(information);
```

Fig 3.1



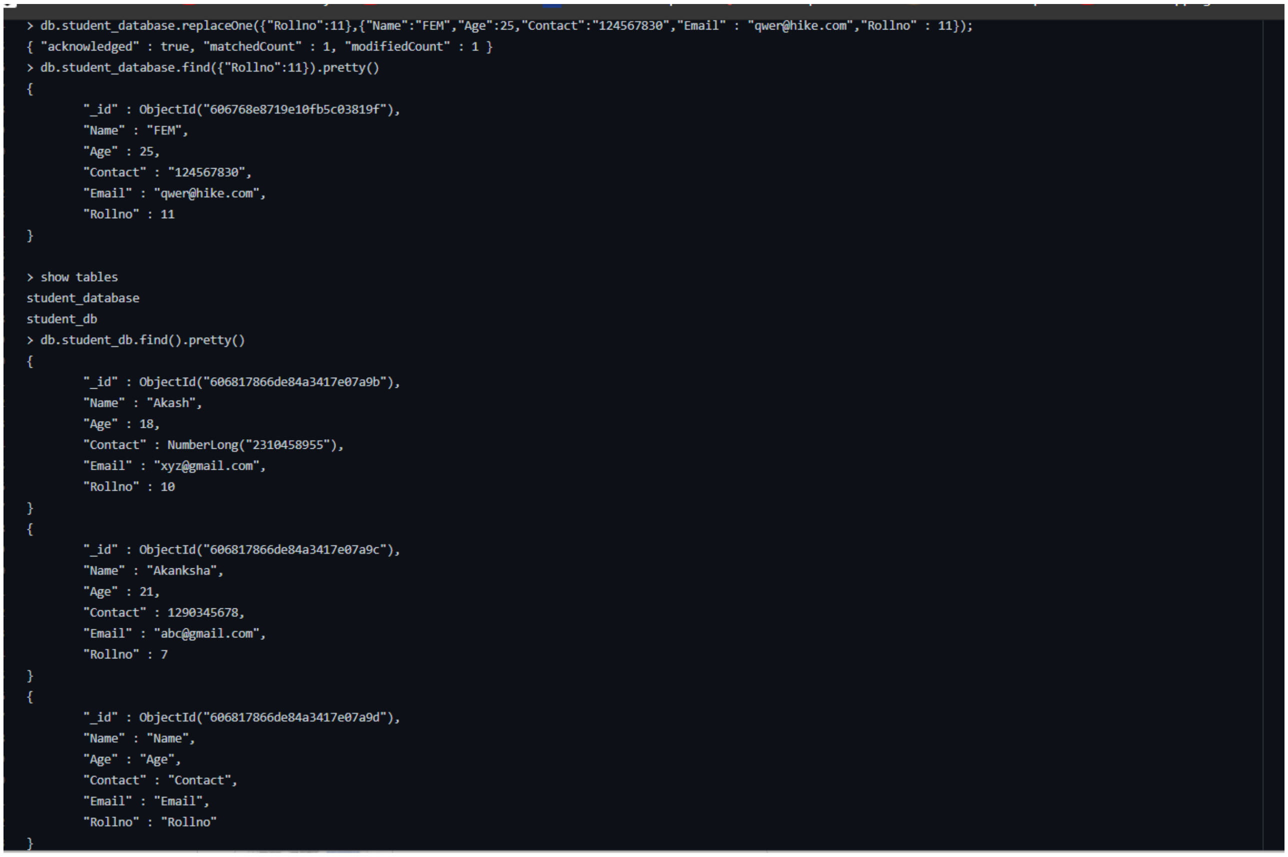
```
> db.student_database.insert(information);
BulkWriteResult({
        "writeErrors" : [ ],
        "writeConcernErrors" : [ ],
        "nInserted" : 4,
        "nUpserted" : 0,
        "nMatched" : 0,
        "nModified" : 0,
        "nRemoved" : 0,
        "upserted" : [ ]
})
> db.student_database.find().pretty()
{
        "_id" : ObjectId("606768e8719e10fb5c03819d"),
        "Name" : "Akanksha",
        "Age" : 21,
        "Contact" : 1290345678,
        "Email" : "abc@gmail.com",
        "Rollno" : 7
}
{
        "_id" : ObjectId("606768e8719e10fb5c03819e"),
        "Name" : "Akash",
        "Age" : 18,
        "Conatct" : 2310458955,
        "Email" : "xyz@yahoo.com",
        "Rollno" : 10
}
{
        "_id" : ObjectId("606768e8719e10fb5c03819f"),
        "Name" : "Mohit",
        "Age" : 25,
        "Conatct" : 124567830,
        "Email" : "qwer@hike.com",
        "Rollno" : 11
}
```

Fig 3.2



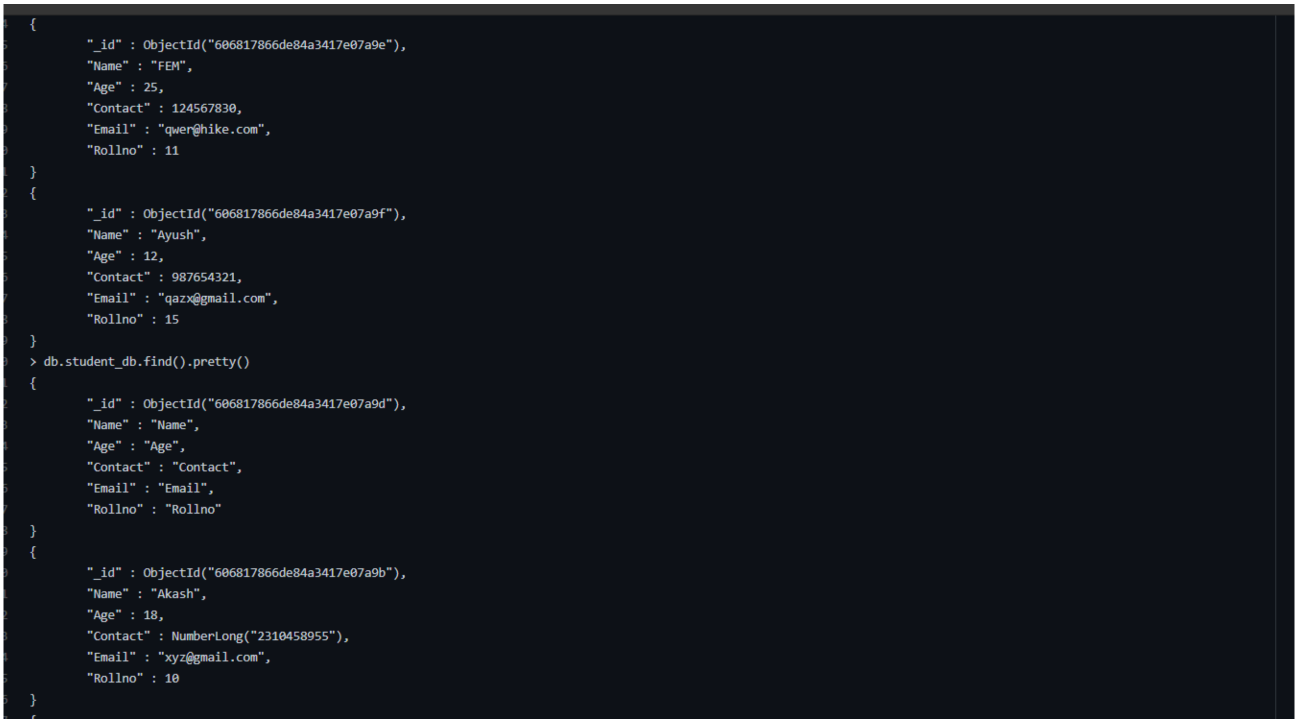
```
{
        "_id" : ObjectId("606768e8719e10fb5c0381a0"),
        "Name" : "Ayush",
        "Age" : 12,
        "Contact" : 987654321,
        "Email" : "qazx@gmail.com",
        "Rollno" : 15
}
> db.student_database.update({"Rollno":10},{$set:{"Email":"xyz@gmail.com"}})
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
> db.student_database.find({"Rollno":10}).pretty()
{
        "_id" : ObjectId("606768e8719e10fb5c03819e"),
        "Name" : "Akash",
        "Age" : 18,
        "Conatct" : 2310458955,
        "Email" : "xyz@gmail.com",
        "Rollno" : 10
}
> db.student_database.update({"Rollno":11},{$set:{"Name":"Piyush"}})
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
> db.student_database.find({"Rollno":11}).pretty()
{
        "_id" : ObjectId("606768e8719e10fb5c03819f"),
        "Name" : "Piyush",
        "Age" : 25,
        "Conatct" : 124567830,
        "Email" : "qwer@hike.com",
        "Rollno" : 11
}
```

Fig 3.3



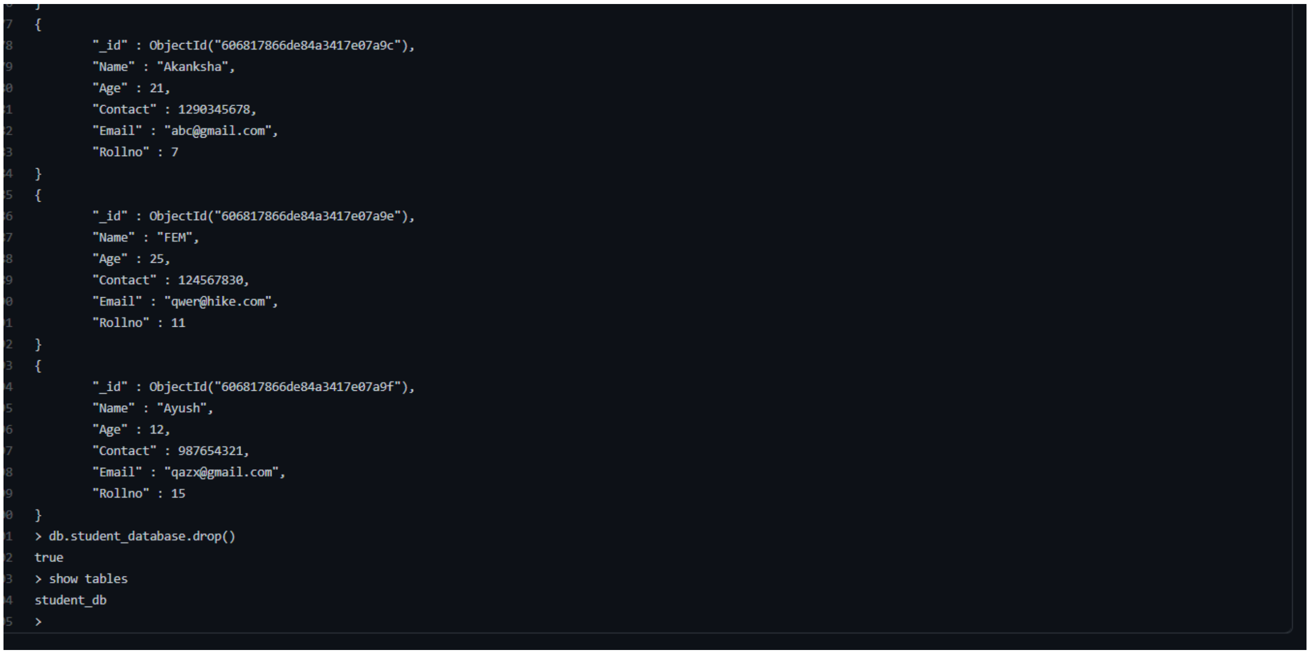
```
> db.student_database.replaceOne({"Rollno":11},{"Name":"FEM","Age":25,"Contact":"124567830","Email" : "qwer@hike.com","Rollno" : 11});
{ "acknowledged" : true, "matchedCount" : 1, "modifiedCount" : 1 }
> db.student_database.find({"Rollno":11}).pretty()
{
        "_id" : ObjectId("606768e8719e10fb5c03819f"),
        "Name" : "FEM",
        "Age" : 25,
        "Contact" : "124567830",
        "Email" : "qwer@hike.com",
        "Rollno" : 11
}

> show tables
student_database
student_db
> db.student_db.find().pretty()
{
        "_id" : ObjectId("606817866de84a3417e07a9b"),
        "Name" : "Akash",
        "Age" : 18,
        "Contact" : NumberLong("2310458955"),
        "Email" : "xyz@gmail.com",
        "Rollno" : 10
}
{
        "_id" : ObjectId("606817866de84a3417e07a9c"),
        "Name" : "Akanksha",
        "Age" : 21,
        "Contact" : 1290345678,
        "Email" : "abc@gmail.com",
        "Rollno" : 7
}
{
        "_id" : ObjectId("606817866de84a3417e07a9d"),
        "Name" : "Name",
        "Age" : "Age",
        "Contact" : "Contact",
        "Email" : "Email",
        "Rollno" : "Rollno"
}
```

Fig 3.4



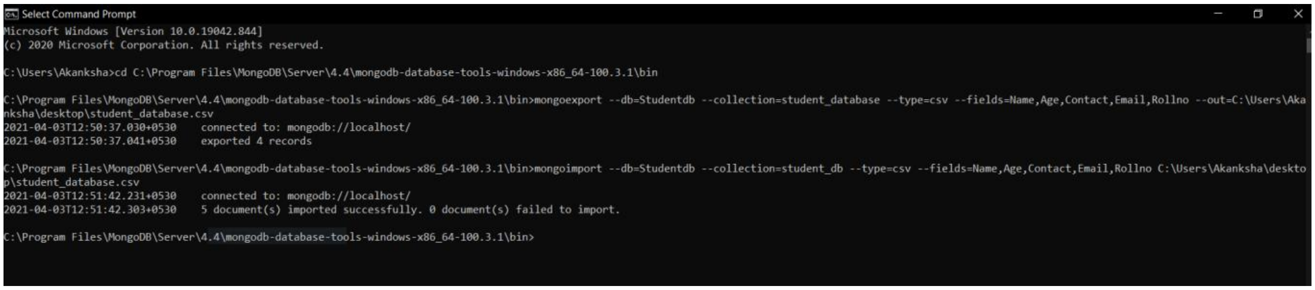
```
{
        "_id" : ObjectId("606817866de84a3417e07a9e"),
        "Name" : "FEM",
        "Age" : 25,
        "Contact" : 124567830,
        "Email" : "qwer@hike.com",
        "Rollno" : 11
}
{
        "_id" : ObjectId("606817866de84a3417e07a9f"),
        "Name" : "Ayush",
        "Age" : 12,
        "Contact" : 987654321,
        "Email" : "qazx@gmail.com",
        "Rollno" : 15
}
> db.student_db.find().pretty()
{
        "_id" : ObjectId("606817866de84a3417e07a9d"),
        "Name" : "Name",
        "Age" : "Age",
        "Contact" : "Contact",
        "Email" : "Email",
        "Rollno" : "Rollno"
}
{
        "_id" : ObjectId("606817866de84a3417e07a9b"),
        "Name" : "Akash",
        "Age" : 18,
        "Contact" : NumberLong("2310458955"),
        "Email" : "xyz@gmail.com",
        "Rollno" : 10
}
```

Fig 3.5



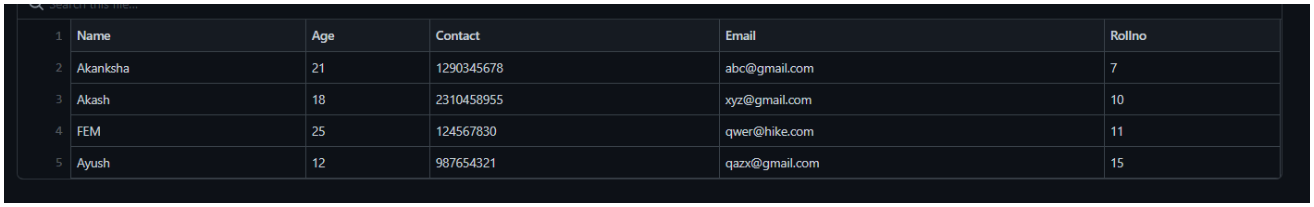
```
{
        "_id" : ObjectId("606817866de84a3417e07a9c"),
        "Name" : "Akanksha",
        "Age" : 21,
        "Contact" : 1290345678,
        "Email" : "abc@gmail.com",
        "Rollno" : 7
}
{
        "_id" : ObjectId("606817866de84a3417e07a9e"),
        "Name" : "FEM",
        "Age" : 25,
        "Contact" : 124567830,
        "Email" : "qwer@hike.com",
        "Rollno" : 11
}
{
        "_id" : ObjectId("606817866de84a3417e07a9f"),
        "Name" : "Ayush",
        "Age" : 12,
        "Contact" : 987654321,
        "Email" : "qazx@gmail.com",
        "Rollno" : 15
}
> db.student_database.drop()
true
> show tables
student_db
>
```

Fig 3.6



```
Microsoft Windows [Version 10.0.19042.844]
(c) 2020 Microsoft Corporation. All rights reserved.

C:\Users\Akanksha>cd C:\Program Files\MongoDB\Server\4.4\mongodb-database-tools-windows-x86_64-100.3.1\bin

C:\Program Files\MongoDB\Server\4.4\mongodb-database-tools-windows-x86_64-100.3.1\bin>mongoexport --db=Studentdb --collection=student_database --type=csv --fields=Name,Age,Contact,Email,Rollno --out=C:\Users\Aka
nksha\desktop\student_database.csv
2021-04-03T12:50:37.030+0530    connected to: mongodb://localhost/
2021-04-03T12:50:37.041+0530    exported 4 records

C:\Program Files\MongoDB\Server\4.4\mongodb-database-tools-windows-x86_64-100.3.1\bin>mongoimport --db=Studentdb --collection=student_db --type=csv --fields=Name,Age,Contact,Email,Rollno C:\Users\Akanksha\deskto
p\student_database.csv
2021-04-03T12:51:42.231+0530    connected to: mongodb://localhost/
2021-04-03T12:51:42.303+0530    5 document(s) imported successfully. 0 document(s) failed to import.

C:\Program Files\MongoDB\Server\4.4\mongodb-database-tools-windows-x86_64-100.3.1\bin>
```

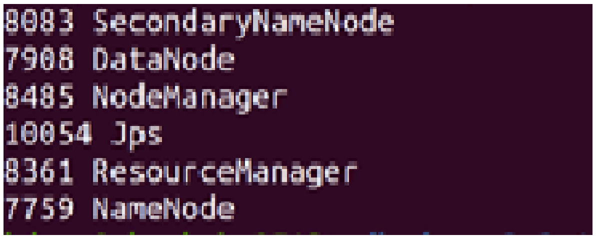
Fig 3.7



Name	Age	Contact	Email	Rollno
Akanksha	21	1290345678	abc@gmail.com	7
Akash	18	2310458955	xyz@gmail.com	10
FEM	25	124567830	qwer@hike.com	11
Ayush	12	987654321	qazx@gmail.com	15

Fig 3.8

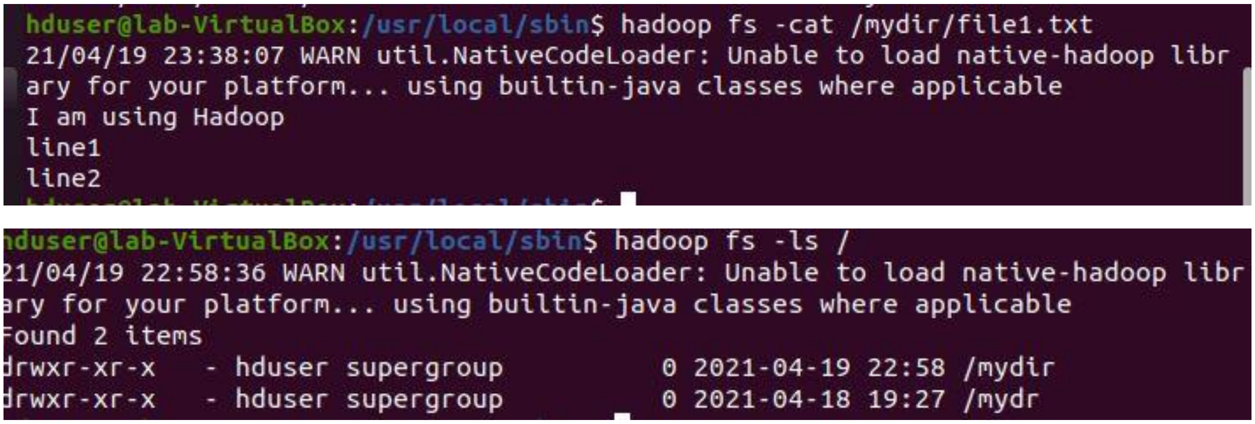
Program 4. Screenshot of Hadoop installed



```
8083 SecondaryNameNode
7908 DataNode
8485 NodeManager
10054 Jps
8361 ResourceManager
7759 NameNode
```

Fig 4.1

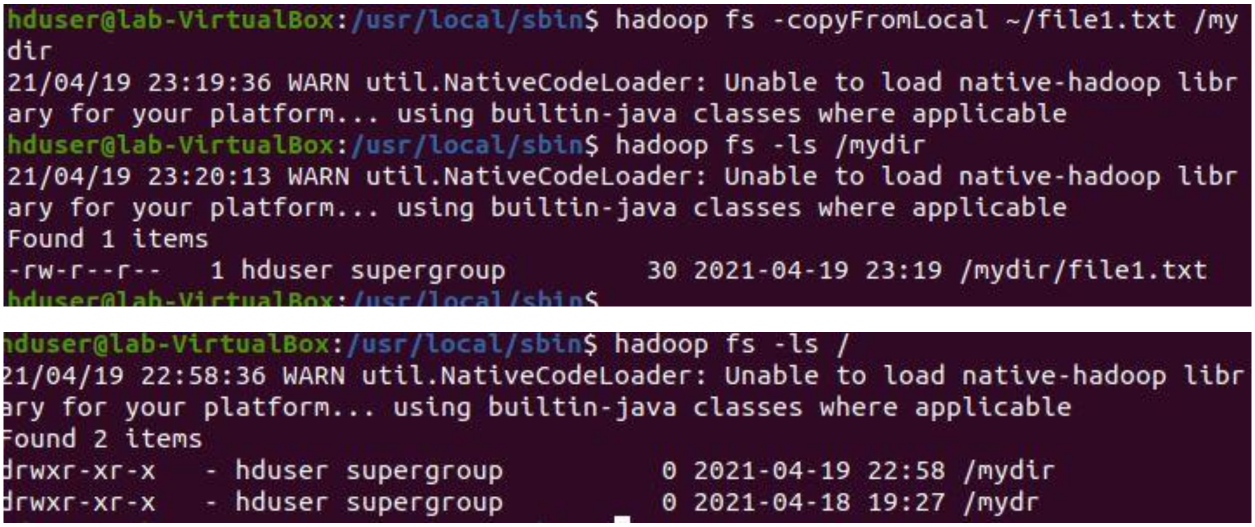
Program 5. Execution of HDFS Commands for interaction with Hadoop Environment. (Minimum 10 commands to be executed).



```
hduser@lab-VirtualBox:/usr/local/sbin$ hadoop fs -cat /mydir/file1.txt
21/04/19 23:38:07 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
I am using Hadoop
line1
line2
```

```
hduser@lab-VirtualBox:/usr/local/sbin$ hadoop fs -ls /
21/04/19 22:58:36 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x   - hduser supergroup          0 2021-04-19 22:58 /mydir
drwxr-xr-x   - hduser supergroup          0 2021-04-18 19:27 /mydr
```

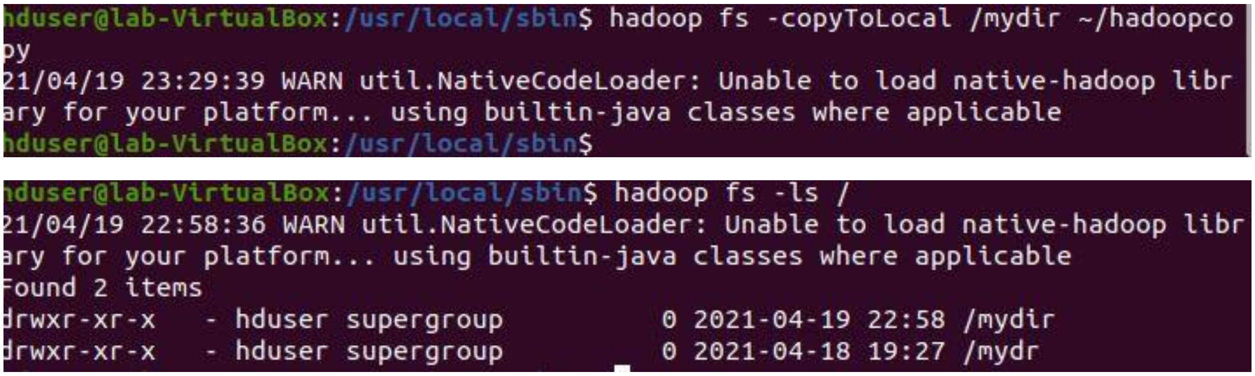
Fig 5.1



```
hduser@lab-VirtualBox:/usr/local/sbin$ hadoop fs -copyFromLocal ~/file1.txt /my
dir
21/04/19 23:19:36 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
hduser@lab-VirtualBox:/usr/local/sbin$ hadoop fs -ls /mydir
21/04/19 23:20:13 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r--   1 hduser supergroup         30 2021-04-19 23:19 /mydir/file1.txt
hduser@lab-VirtualBox:/usr/local/sbin$
```

```
hduser@lab-VirtualBox:/usr/local/sbin$ hadoop fs -ls /
21/04/19 22:58:36 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x   - hduser supergroup          0 2021-04-19 22:58 /mydir
drwxr-xr-x   - hduser supergroup          0 2021-04-18 19:27 /mydr
```

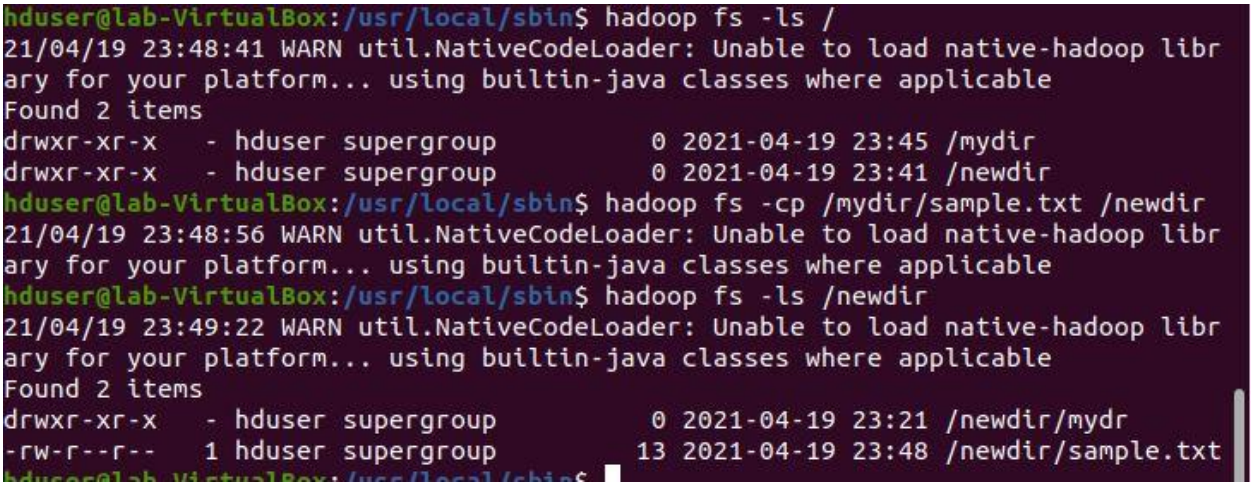
Fig 5.2



```
hduser@lab-VirtualBox:/usr/local/sbin$ hadoop fs -copyToLocal /mydir ~/hadoopco
py
21/04/19 23:29:39 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
hduser@lab-VirtualBox:/usr/local/sbin$
```

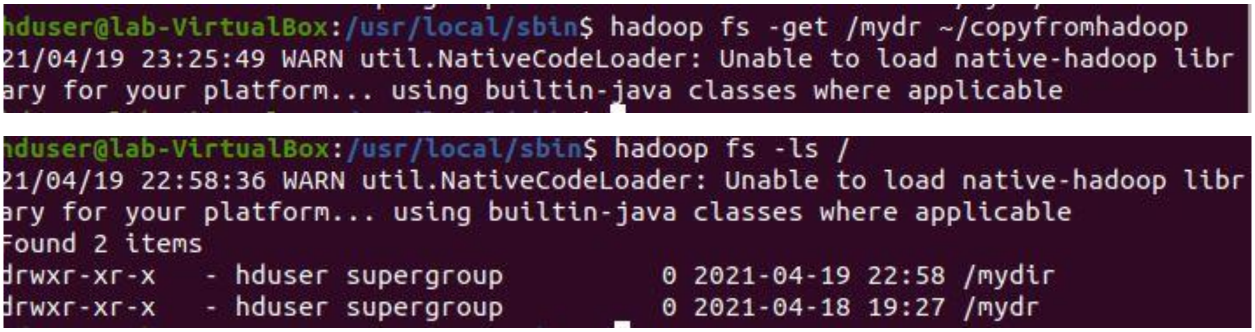
```
hduser@lab-VirtualBox:/usr/local/sbin$ hadoop fs -ls /
21/04/19 22:58:36 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x   - hduser supergroup          0 2021-04-19 22:58 /mydir
drwxr-xr-x   - hduser supergroup          0 2021-04-18 19:27 /mydr
```

Fig 5.3



```
hduser@lab-VirtualBox:/usr/local/sbin$ hadoop fs -ls /
21/04/19 23:48:41 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x   - hduser supergroup          0 2021-04-19 23:45 /mydir
drwxr-xr-x   - hduser supergroup          0 2021-04-19 23:41 /newdir
hduser@lab-VirtualBox:/usr/local/sbin$ hadoop fs -cp /mydir/sample.txt /newdir
21/04/19 23:48:56 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
hduser@lab-VirtualBox:/usr/local/sbin$ hadoop fs -ls /newdir
21/04/19 23:49:22 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x   - hduser supergroup          0 2021-04-19 23:21 /newdir/mydr
-rw-r--r--   1 hduser supergroup         13 2021-04-19 23:48 /newdir/sample.txt
hduser@lab-VirtualBox:/usr/local/sbin$
```

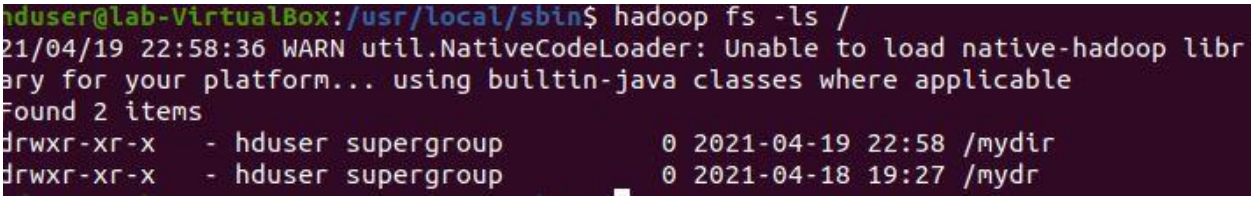
Fig 5.4



```
hduser@lab-VirtualBox:/usr/local/sbin$ hadoop fs -get /mydr ~/copyfromhadoop
21/04/19 23:25:49 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
```

```
hduser@lab-VirtualBox:/usr/local/sbin$ hadoop fs -ls /
21/04/19 22:58:36 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x   - hduser supergroup          0 2021-04-19 22:58 /mydir
drwxr-xr-x   - hduser supergroup          0 2021-04-18 19:27 /mydr
```

Fig 5.5



```
hduser@lab-VirtualBox:/usr/local/sbin$ hadoop fs -ls /
21/04/19 22:58:36 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
Found 2 items
drwxr-xr-x   - hduser supergroup          0 2021-04-19 22:58 /mydir
drwxr-xr-x   - hduser supergroup          0 2021-04-18 19:27 /mydr
```

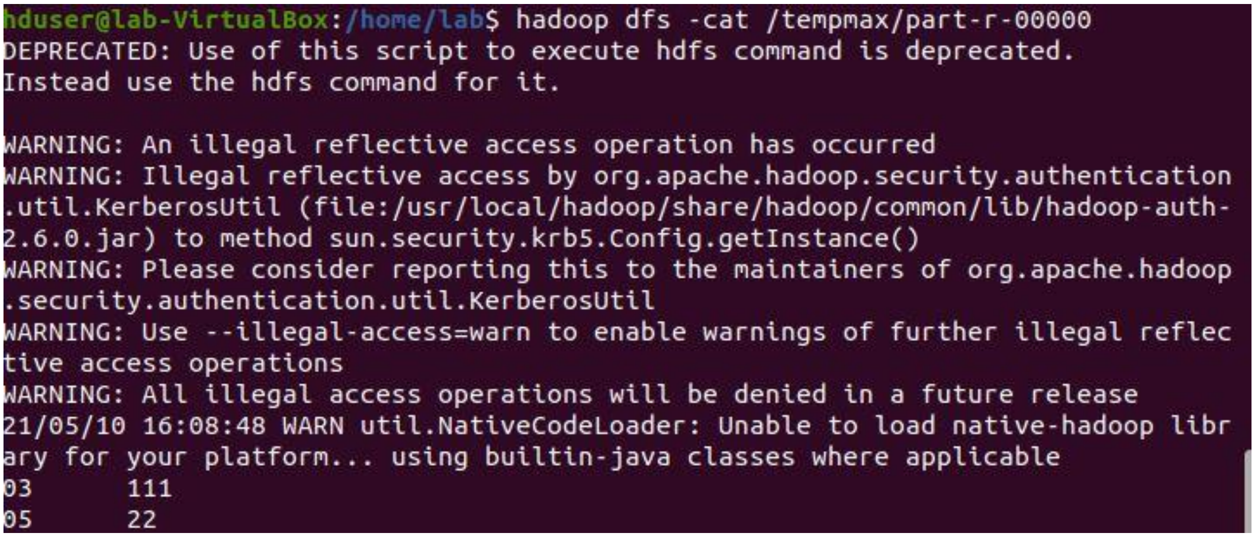
Fig 5.6

Program 6. From the following link extract the weather data

https://github.com/tomwhite/hadoop-book/tree/master/input/ncdc/all .Create a Map Reduce program to

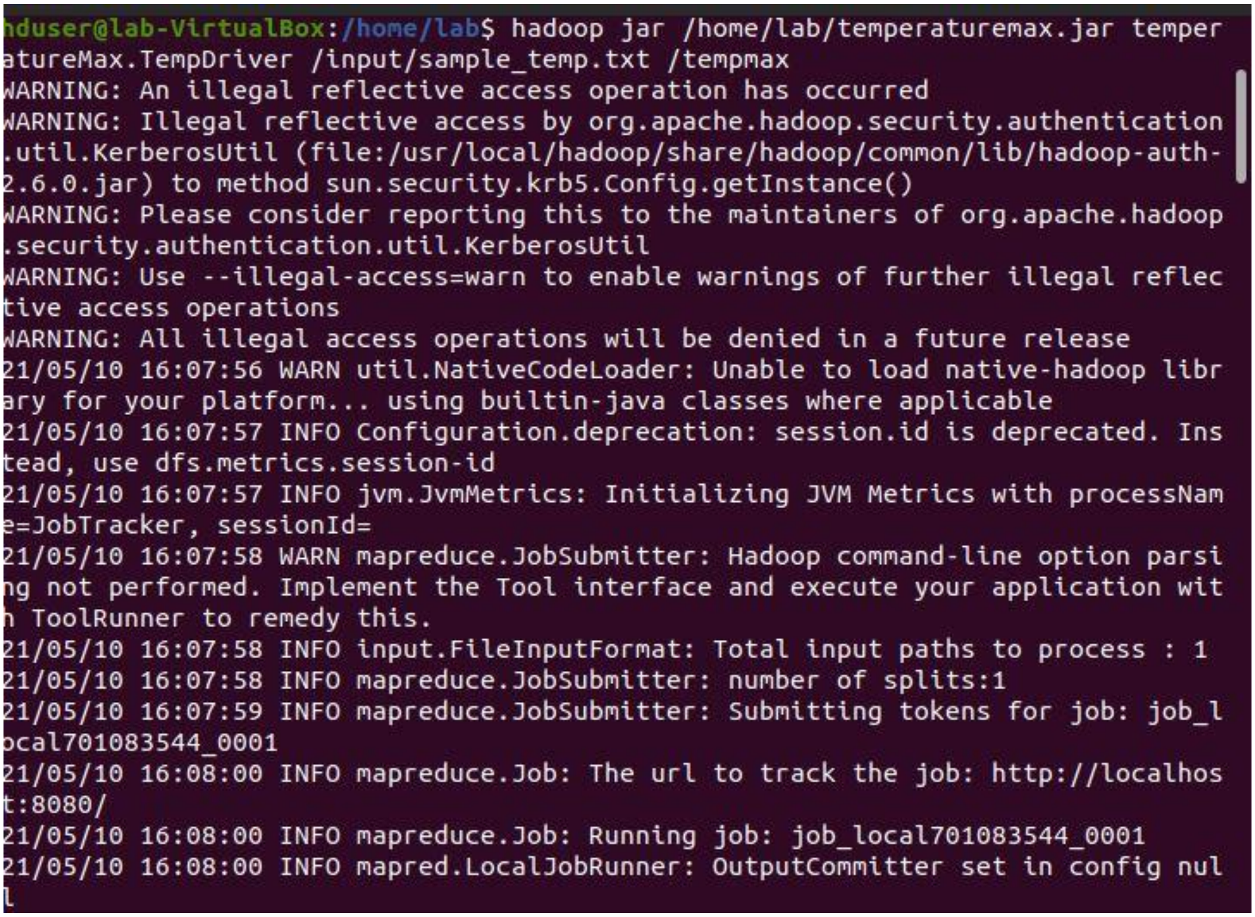
i) find average temperature for each year from NCDC data set.

ii) find the mean max temperature for every month.



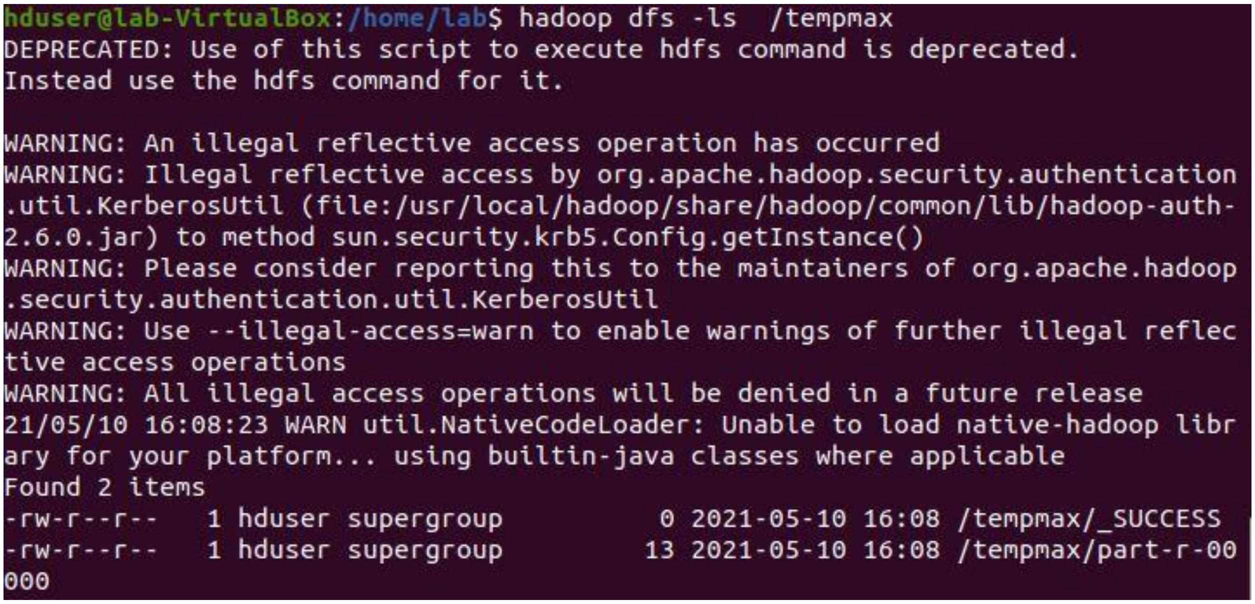
```
hduser@lab-VirtualBox:/home/lab$ hadoop dfs -cat /tempmax/part-r-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication
.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-
2.6.0.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop
.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflec
tive access operations
WARNING: All illegal access operations will be denied in a future release
21/05/10 16:08:48 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
03      111
05      22
```

Fig 6.1.1



```
hduser@lab-VirtualBox:/home/lab$ hadoop jar /home/lab/temperaturemax.jar temper
atureMax.TempDriver /input/sample_temp.txt /tempmax
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication
.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-
2.6.0.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop
.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflec
tive access operations
WARNING: All illegal access operations will be denied in a future release
21/05/10 16:07:56 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
21/05/10 16:07:57 INFO Configuration.deprecation: session.id is deprecated. Ins
tead, use dfs.metrics.session-id
21/05/10 16:07:57 INFO jvm.JvmMetrics: Initializing JVM Metrics with processNam
e=JobTracker, sessionId=
21/05/10 16:07:58 WARN mapreduce.JobSubmitter: Hadoop command-line option parsi
ng not performed. Implement the Tool interface and execute your application wit
h ToolRunner to remedy this.
21/05/10 16:07:58 INFO input.FileInputFormat: Total input paths to process : 1
21/05/10 16:07:58 INFO mapreduce.JobSubmitter: number of splits:1
21/05/10 16:07:59 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_l
ocal701083544_0001
21/05/10 16:08:00 INFO mapreduce.Job: The url to track the job: http://localhos
t:8080/
21/05/10 16:08:00 INFO mapreduce.Job: Running job: job_local701083544_0001
21/05/10 16:08:00 INFO mapred.LocalJobRunner: OutputCommitter set in config nul
l
```

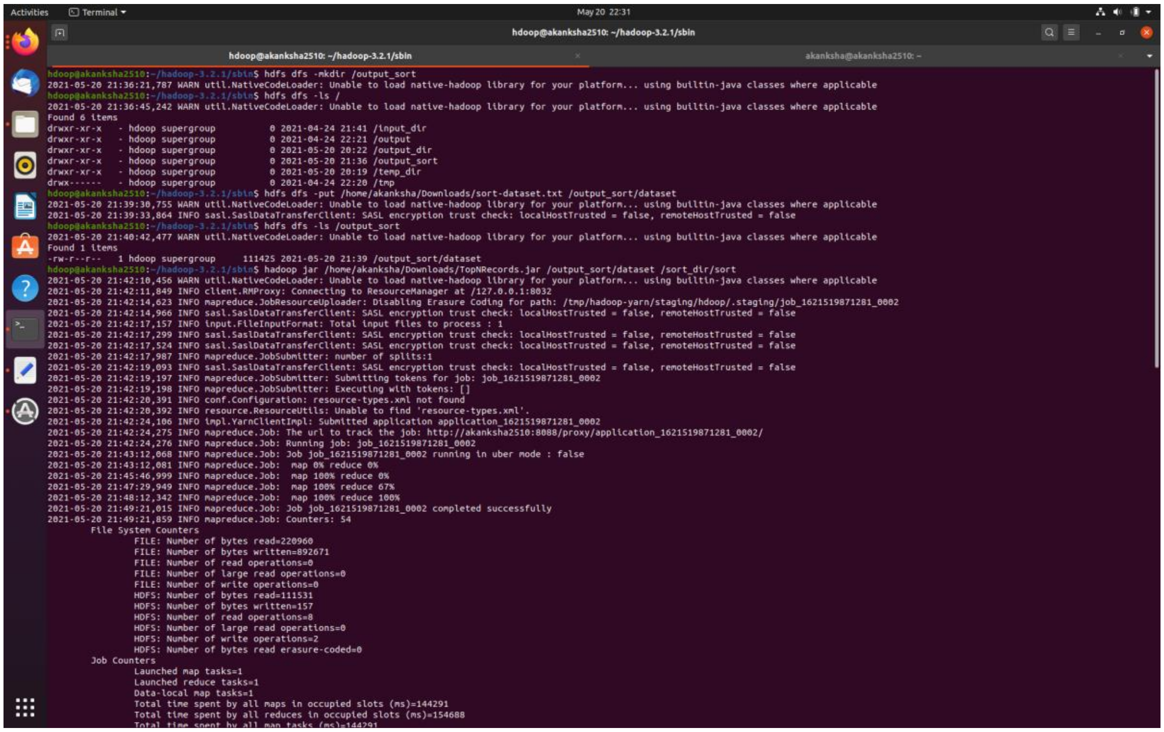
Fig 6.1.2



```
hduser@lab-VirtualBox:/home/lab$ hadoop dfs -ls  /tempmax
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication
.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-
2.6.0.jar) to method sun.security.krb5.Config.getInstance()
WARNING: Please consider reporting this to the maintainers of org.apache.hadoop
.security.authentication.util.KerberosUtil
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflec
tive access operations
WARNING: All illegal access operations will be denied in a future release
21/05/10 16:08:23 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
ary for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 hduser supergroup          0 2021-05-10 16:08 /tempmax/_SUCCESS
-rw-r--r--   1 hduser supergroup         13 2021-05-10 16:08 /tempmax/part-r-00
000
```

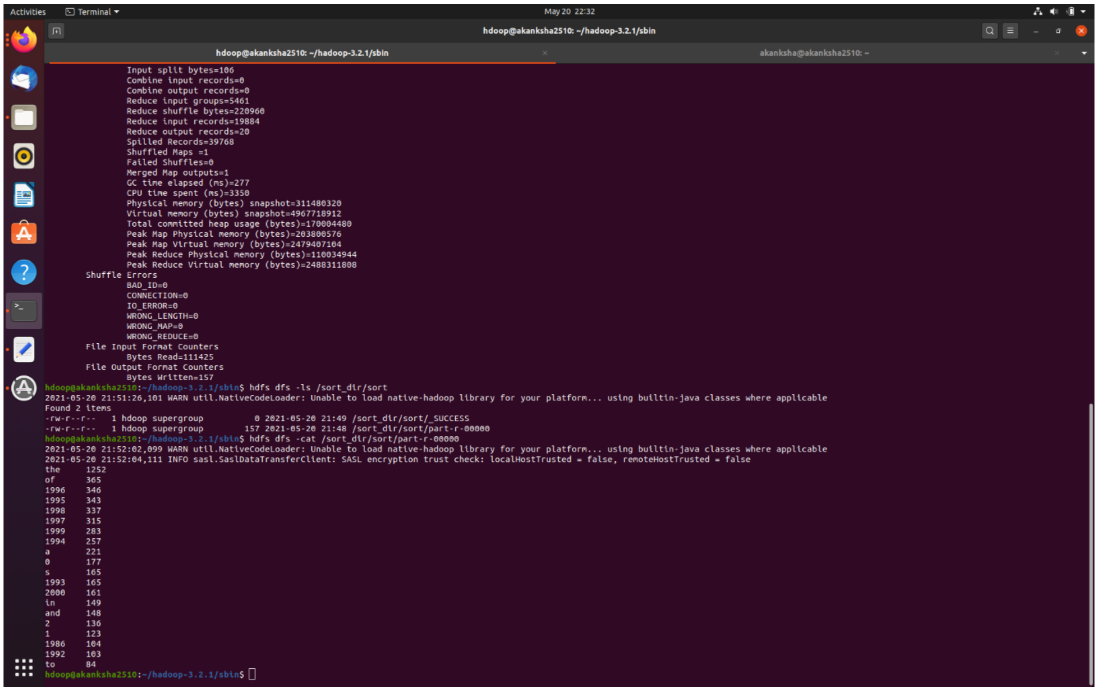
Fig 6.2

Program 7. For a given Text file, create a Map Reduce program to sort the content in an alphabetic order listing only top 10 maximum occurrences of words.


Activities
Terminal
May 20 22:31
hdoop@akanksha2510: ~/hadoop-3.2.1/sbin
hdoop@akanksha2510: ~/hadoop-3.2.1/sbin
akanksha@akanksha2510: ~

```
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ hdfs dfs -mkdir /output_sort
2021-05-20 21:36:21,787 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ hdfs dfs -ls /
2021-05-20 21:36:45,242 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 6 items
drwxr-xr-x   - hdoop supergroup          0 2021-04-24 21:41 /input_dir
drwxr-xr-x   - hdoop supergroup          0 2021-04-24 22:21 /output
drwxr-xr-x   - hdoop supergroup          0 2021-05-20 20:22 /output_dir
drwxr-xr-x   - hdoop supergroup          0 2021-05-20 21:36 /output_sort
drwxr-xr-x   - hdoop supergroup          0 2021-05-20 20:19 /temp_dir
drwx------   - hdoop supergroup          0 2021-04-24 22:20 /tmp
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ hdfs dfs -put /home/akanksha/Downloads/sort-dataset.txt /output_sort/dataset
2021-05-20 21:39:30,755 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2021-05-20 21:39:33,864 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ hdfs dfs -ls /output_sort
2021-05-20 21:40:42,477 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r--   1 hdoop supergroup     111425 2021-05-20 21:39 /output_sort/dataset
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ hadoop jar /home/akanksha/Downloads/TopNRecords.jar /output_sort/dataset /sort_dir/sort
2021-05-20 21:42:10,456 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2021-05-20 21:42:11,849 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-05-20 21:42:14,623 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_1621519871281_0002
2021-05-20 21:42:14,966 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-05-20 21:42:17,157 INFO input.FileInputFormat: Total input files to process : 1
2021-05-20 21:42:17,299 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-05-20 21:42:17,524 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-05-20 21:42:17,987 INFO mapreduce.JobSubmitter: number of splits:1
2021-05-20 21:42:19,093 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-05-20 21:42:19,197 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1621519871281_0002
2021-05-20 21:42:19,198 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-05-20 21:42:20,391 INFO conf.Configuration: resource-types.xml not found
2021-05-20 21:42:20,392 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-05-20 21:42:24,106 INFO impl.YarnClientImpl: Submitted application application_1621519871281_0002
2021-05-20 21:42:24,275 INFO mapreduce.Job: The url to track the job: http://akanksha2510:8088/proxy/application_1621519871281_0002/
2021-05-20 21:42:24,276 INFO mapreduce.Job: Running job: job_1621519871281_0002
2021-05-20 21:43:12,068 INFO mapreduce.Job: Job job_1621519871281_0002 running in uber mode : false
2021-05-20 21:43:12,081 INFO mapreduce.Job:  map 0% reduce 0%
2021-05-20 21:45:46,999 INFO mapreduce.Job:  map 100% reduce 0%
2021-05-20 21:47:29,949 INFO mapreduce.Job:  map 100% reduce 67%
2021-05-20 21:48:12,342 INFO mapreduce.Job:  map 100% reduce 100%
2021-05-20 21:49:21,015 INFO mapreduce.Job: Job job_1621519871281_0002 completed successfully
2021-05-20 21:49:21,859 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=220960
                FILE: Number of bytes written=892671
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=111531
                HDFS: Number of bytes written=157
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=144291
                Total time spent by all reduces in occupied slots (ms)=154688
                Total time spent by all map tasks (ms)=144291
```

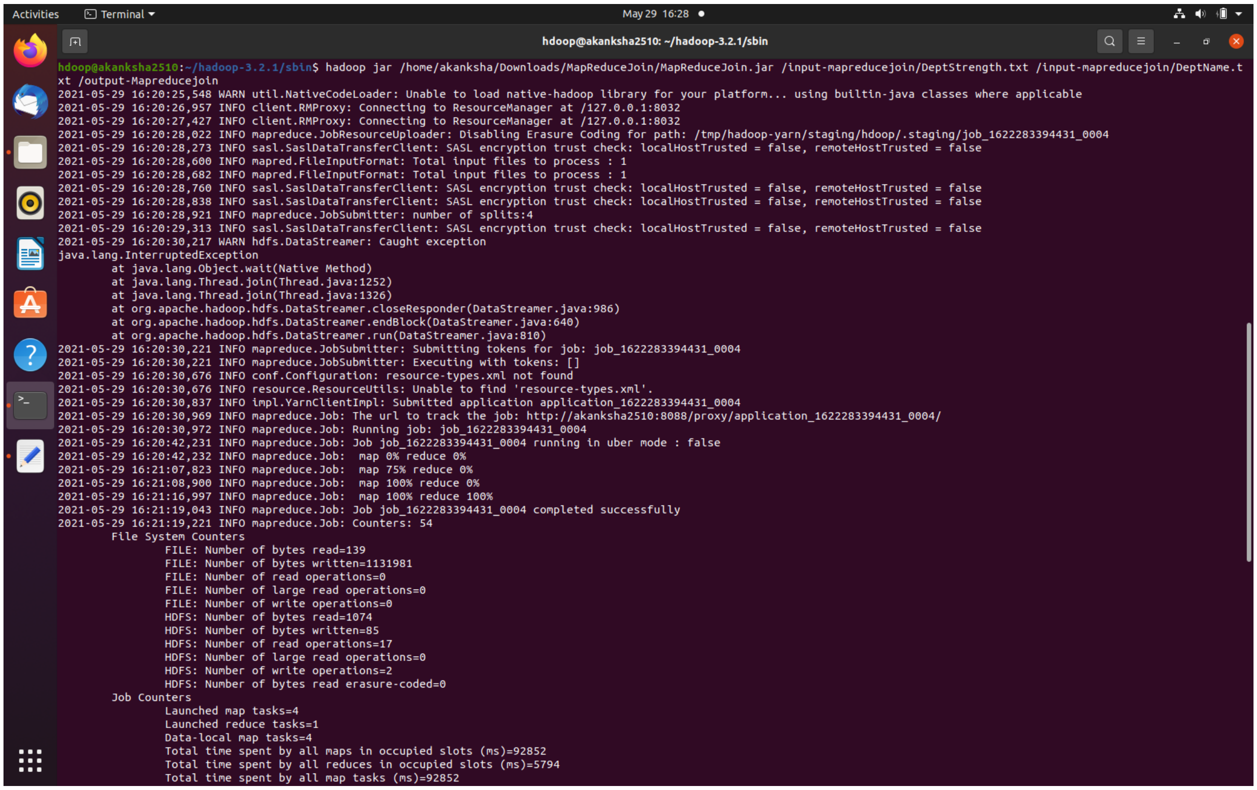
Fig 7.1


Activities
Terminal
May 20 22:32
hdoop@akanksha2510: ~/hadoop-3.2.1/sbin
hdoop@akanksha2510: ~/hadoop-3.2.1/sbin
akanksha@akanksha2510: ~

```
                Input split bytes=106
                Combine input records=0
                Combine output records=0
                Reduce input groups=5461
                Reduce shuffle bytes=220960
                Reduce input records=19884
                Reduce output records=20
                Spilled Records=39768
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=277
                CPU time spent (ms)=3350
                Physical memory (bytes) snapshot=311480320
                Virtual memory (bytes) snapshot=4967718912
                Total committed heap usage (bytes)=170004480
                Peak Map Physical memory (bytes)=203800576
                Peak Map Virtual memory (bytes)=2479407104
                Peak Reduce Physical memory (bytes)=110034944
                Peak Reduce Virtual memory (bytes)=2488311808
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=111425
        File Output Format Counters
                Bytes Written=157
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ hdfs dfs -ls /sort_dir/sort
2021-05-20 21:51:26,101 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 hdoop supergroup          0 2021-05-20 21:49 /sort_dir/sort/_SUCCESS
-rw-r--r--   1 hdoop supergroup        157 2021-05-20 21:48 /sort_dir/sort/part-r-00000
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ hdfs dfs -cat /sort_dir/sort/part-r-00000
2021-05-20 21:52:02,099 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2021-05-20 21:52:04,111 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
the     1252
of      365
1996    346
1995    343
1998    337
1997    315
1999    283
1994    257
a       221
0       177
s       165
1993    165
2000    161
in      149
and     148
2       136
1       123
1986    104
1992    103
to      84
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ 
```

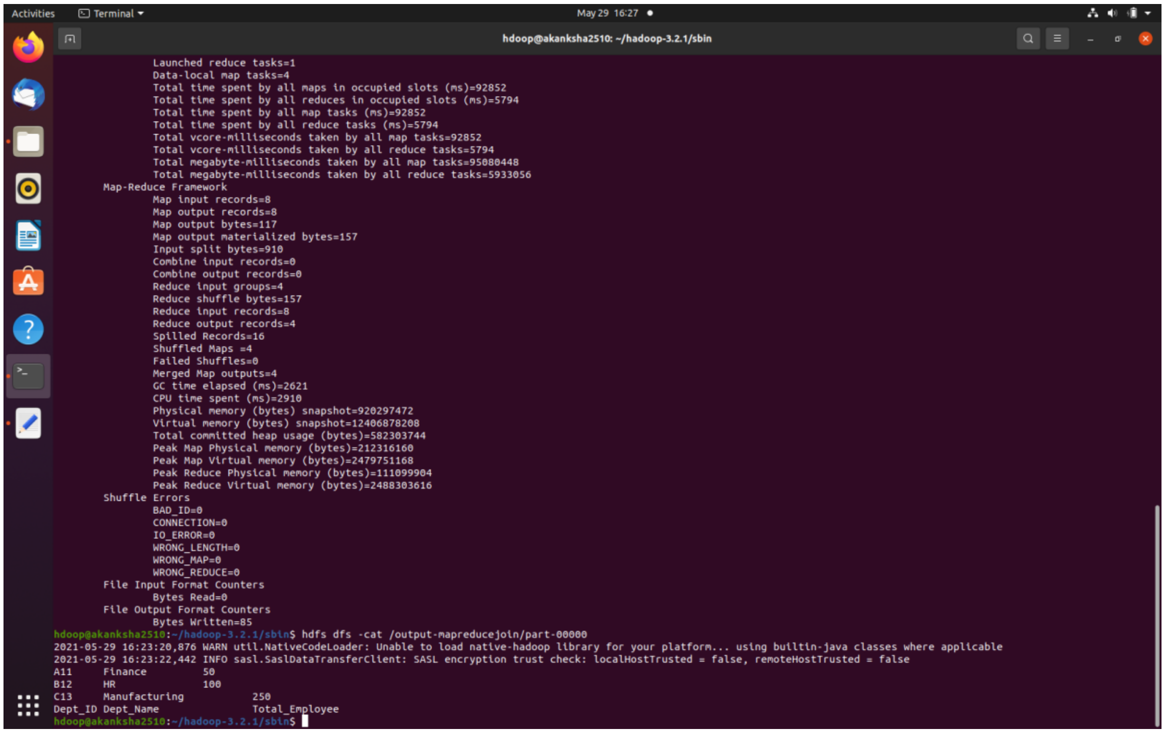
Fig 7.2

Program 8. Create a Map Reduce program to combine information from the users file along with Information from the posts file by using the concept of join and display user_id, Reputation and Score**.**



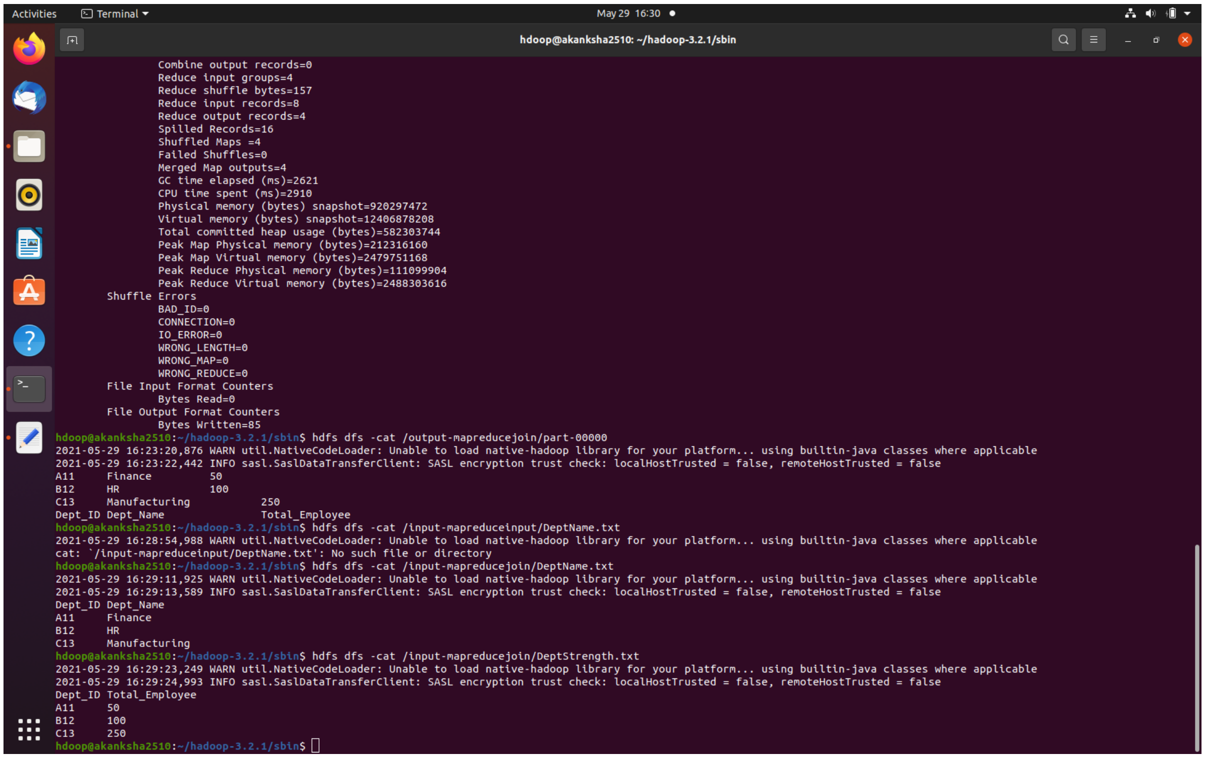
```
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ hadoop jar /home/akanksha/Downloads/MapReduceJoin/MapReduceJoin.jar /input-mapreducejoin/DeptStrength.txt /input-mapreducejoin/DeptName.t
xt /output-Mapreducejoin
2021-05-29 16:20:25,548 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2021-05-29 16:20:26,957 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-05-29 16:20:27,427 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2021-05-29 16:20:28,022 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.staging/job_1622283394431_0004
2021-05-29 16:20:28,273 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-05-29 16:20:28,600 INFO mapred.FileInputFormat: Total input files to process : 1
2021-05-29 16:20:28,682 INFO mapred.FileInputFormat: Total input files to process : 1
2021-05-29 16:20:28,760 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-05-29 16:20:28,838 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-05-29 16:20:28,921 INFO mapreduce.JobSubmitter: number of splits:4
2021-05-29 16:20:29,313 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
2021-05-29 16:20:30,217 WARN hdfs.DataStreamer: Caught exception
java.lang.InterruptedException
        at java.lang.Object.wait(Native Method)
        at java.lang.Thread.join(Thread.java:1252)
        at java.lang.Thread.join(Thread.java:1326)
        at org.apache.hadoop.hdfs.DataStreamer.closeResponder(DataStreamer.java:986)
        at org.apache.hadoop.hdfs.DataStreamer.endBlock(DataStreamer.java:640)
        at org.apache.hadoop.hdfs.DataStreamer.run(DataStreamer.java:810)
2021-05-29 16:20:30,221 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1622283394431_0004
2021-05-29 16:20:30,221 INFO mapreduce.JobSubmitter: Executing with tokens: []
2021-05-29 16:20:30,676 INFO conf.Configuration: resource-types.xml not found
2021-05-29 16:20:30,676 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2021-05-29 16:20:30,837 INFO impl.YarnClientImpl: Submitted application application_1622283394431_0004
2021-05-29 16:20:30,969 INFO mapreduce.Job: The url to track the job: http://akanksha2510:8088/proxy/application_1622283394431_0004/
2021-05-29 16:20:30,972 INFO mapreduce.Job: Running job: job_1622283394431_0004
2021-05-29 16:20:42,231 INFO mapreduce.Job: Job job_1622283394431_0004 running in uber mode : false
2021-05-29 16:20:42,232 INFO mapreduce.Job:  map 0% reduce 0%
2021-05-29 16:21:07,823 INFO mapreduce.Job:  map 75% reduce 0%
2021-05-29 16:21:08,900 INFO mapreduce.Job:  map 100% reduce 0%
2021-05-29 16:21:16,997 INFO mapreduce.Job:  map 100% reduce 100%
2021-05-29 16:21:19,043 INFO mapreduce.Job: Job job_1622283394431_0004 completed successfully
2021-05-29 16:21:19,221 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=139
                FILE: Number of bytes written=1131981
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1074
                HDFS: Number of bytes written=85
                HDFS: Number of read operations=17
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=4
                Launched reduce tasks=1
                Data-local map tasks=4
                Total time spent by all maps in occupied slots (ms)=92852
                Total time spent by all reduces in occupied slots (ms)=5794
                Total time spent by all map tasks (ms)=92852
```

Fig 8.1



```
                Launched reduce tasks=1
                Data-local map tasks=4
                Total time spent by all maps in occupied slots (ms)=92852
                Total time spent by all reduces in occupied slots (ms)=5794
                Total time spent by all map tasks (ms)=92852
                Total time spent by all reduce tasks (ms)=5794
                Total vcore-milliseconds taken by all map tasks=92852
                Total vcore-milliseconds taken by all reduce tasks=5794
                Total megabyte-milliseconds taken by all map tasks=95080448
                Total megabyte-milliseconds taken by all reduce tasks=5933056
        Map-Reduce Framework
                Map input records=8
                Map output records=8
                Map output bytes=117
                Map output materialized bytes=157
                Input split bytes=910
                Combine input records=0
                Combine output records=0
                Reduce input groups=4
                Reduce shuffle bytes=157
                Reduce input records=8
                Reduce output records=4
                Spilled Records=16
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=2621
                CPU time spent (ms)=2910
                Physical memory (bytes) snapshot=920297472
                Virtual memory (bytes) snapshot=12406878208
                Total committed heap usage (bytes)=582303744
                Peak Map Physical memory (bytes)=212316160
                Peak Map Virtual memory (bytes)=2479751168
                Peak Reduce Physical memory (bytes)=111099904
                Peak Reduce Virtual memory (bytes)=2488303616
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=85
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ hdfs dfs -cat /output-mapreducejoin/part-00000
2021-05-29 16:23:20,876 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2021-05-29 16:23:22,442 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
A11     Finance         50
B12     HR              100
C13     Manufacturing           250
Dept_ID Dept_Name               Total_Employee
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ 
```

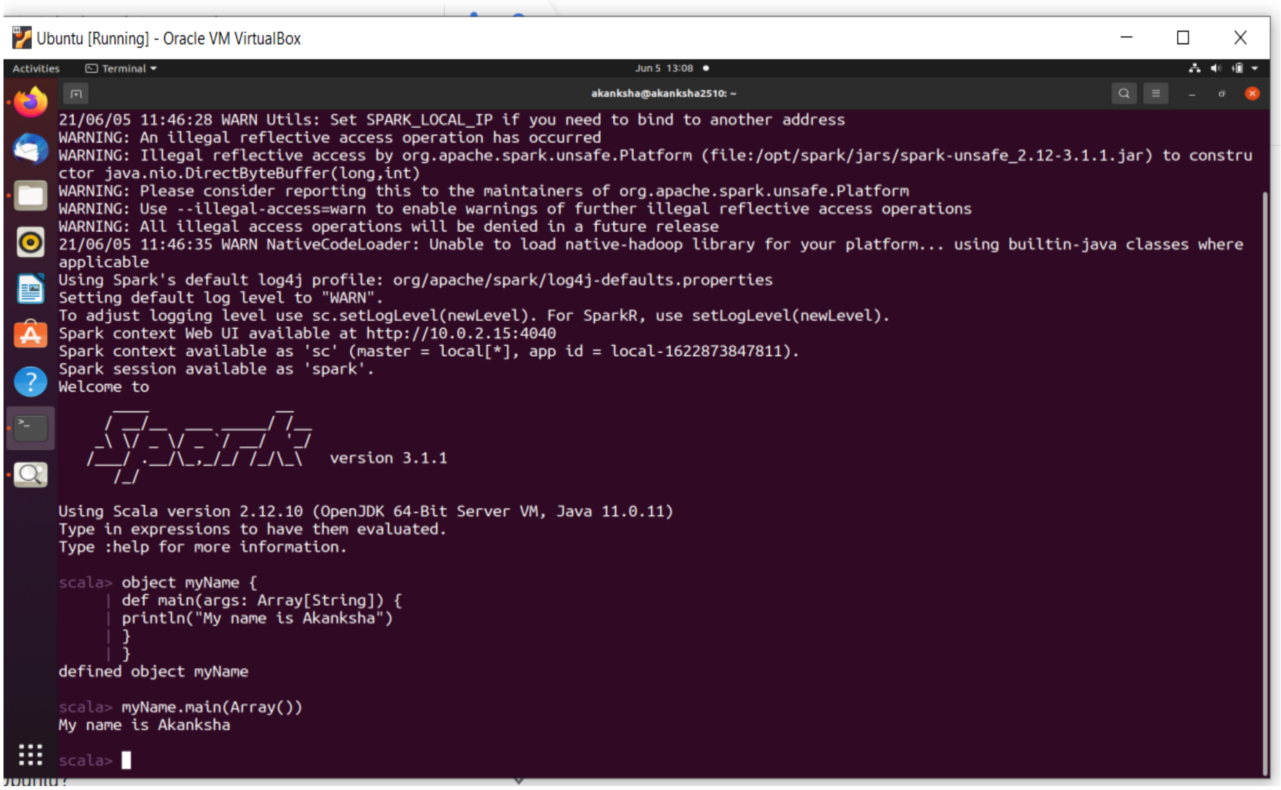
Fig 8.2



```
                Combine output records=0
                Reduce input groups=4
                Reduce shuffle bytes=157
                Reduce input records=8
                Reduce output records=4
                Spilled Records=16
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=2621
                CPU time spent (ms)=2910
                Physical memory (bytes) snapshot=920297472
                Virtual memory (bytes) snapshot=12406878208
                Total committed heap usage (bytes)=582303744
                Peak Map Physical memory (bytes)=212316160
                Peak Map Virtual memory (bytes)=2479751168
                Peak Reduce Physical memory (bytes)=111099904
                Peak Reduce Virtual memory (bytes)=2488303616
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=85
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ hdfs dfs -cat /output-mapreducejoin/part-00000
2021-05-29 16:23:20,876 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2021-05-29 16:23:22,442 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
A11     Finance         50
B12     HR              100
C13     Manufacturing           250
Dept_ID Dept_Name               Total_Employee
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ hdfs dfs -cat /input-mapreduceinput/DeptName.txt
2021-05-29 16:28:54,988 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
cat: `/input-mapreduceinput/DeptName.txt': No such file or directory
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ hdfs dfs -cat /input-mapreducejoin/DeptName.txt
2021-05-29 16:29:11,925 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2021-05-29 16:29:13,589 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Dept_ID Dept_Name
A11     Finance
B12     HR
C13     Manufacturing
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ hdfs dfs -cat /input-mapreducejoin/DeptStrength.txt
2021-05-29 16:29:23,249 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2021-05-29 16:29:24,993 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted = false
Dept_ID Total_Employee
A11     50
B12     100
C13     250
hdoop@akanksha2510:~/hadoop-3.2.1/sbin$ 
```

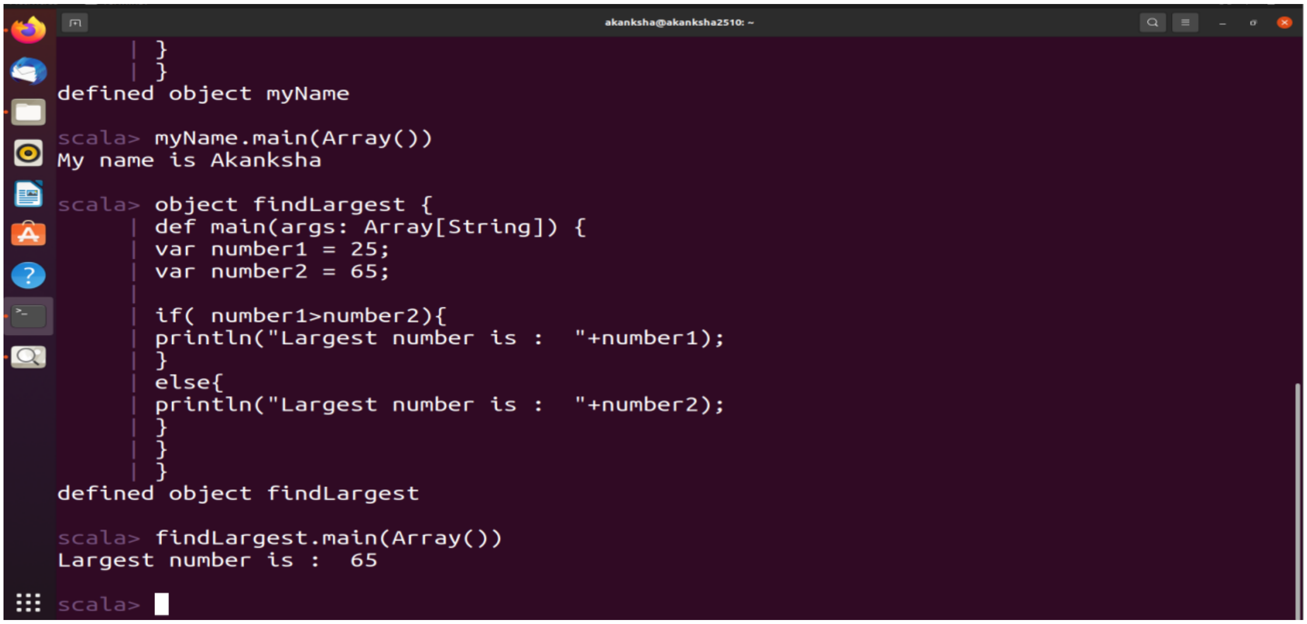
Fig 8.3

Program 9. Scala programs and Screenshot of Spark Installed



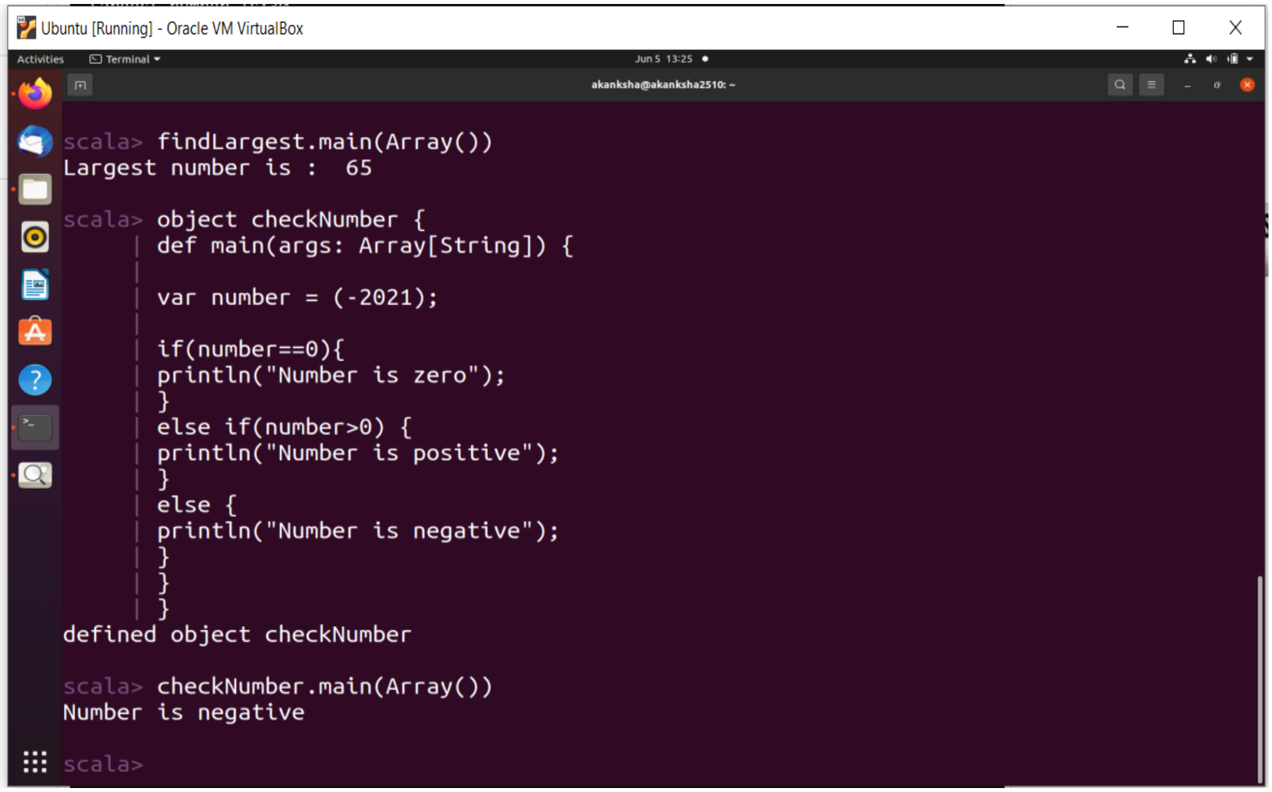
```
21/06/05 11:46:28 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark/jars/spark-unsafe_2.12-3.1.1.jar) to constru
ctor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
21/06/05 11:46:35 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where
applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://10.0.2.15:4040
Spark context available as 'sc' (master = local[*], app id = local-1622873847811).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.1.1
      /_/

Using Scala version 2.12.10 (OpenJDK 64-Bit Server VM, Java 11.0.11)
Type in expressions to have them evaluated.
Type :help for more information.

scala> object myName {
     | def main(args: Array[String]) {
     | println("My name is Akanksha")
     | }
     | }
defined object myName

scala> myName.main(Array())
My name is Akanksha

scala> 
```

Fig 9.1



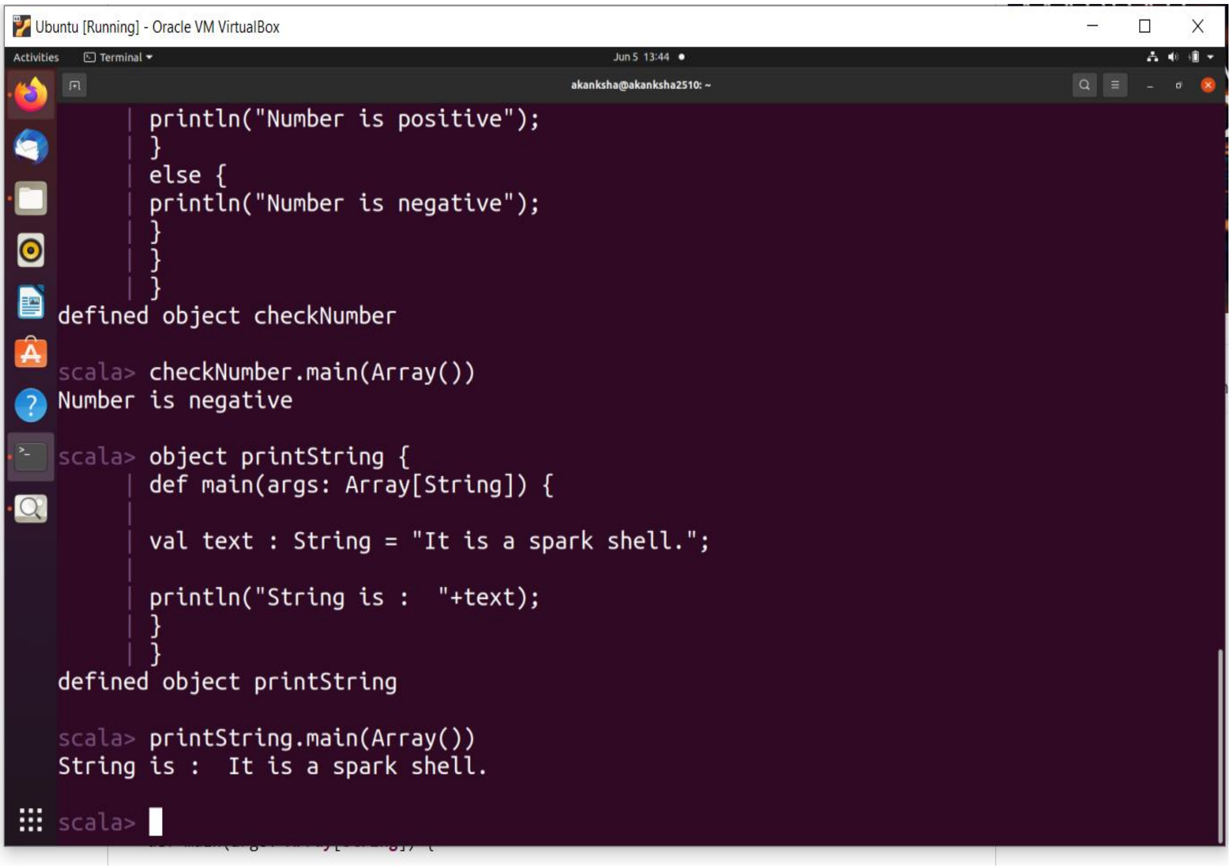
```
     | }
     | }
defined object myName

scala> myName.main(Array())
My name is Akanksha

scala> object findLargest {
     | def main(args: Array[String]) {
     | var number1 = 25;
     | var number2 = 65;
     |
     | if( number1>number2){
     | println("Largest number is :  "+number1);
     | }
     | else{
     | println("Largest number is :  "+number2);
     | }
     | }
     | }
defined object findLargest

scala> findLargest.main(Array())
Largest number is :  65

scala> 
```

Fig 9.2



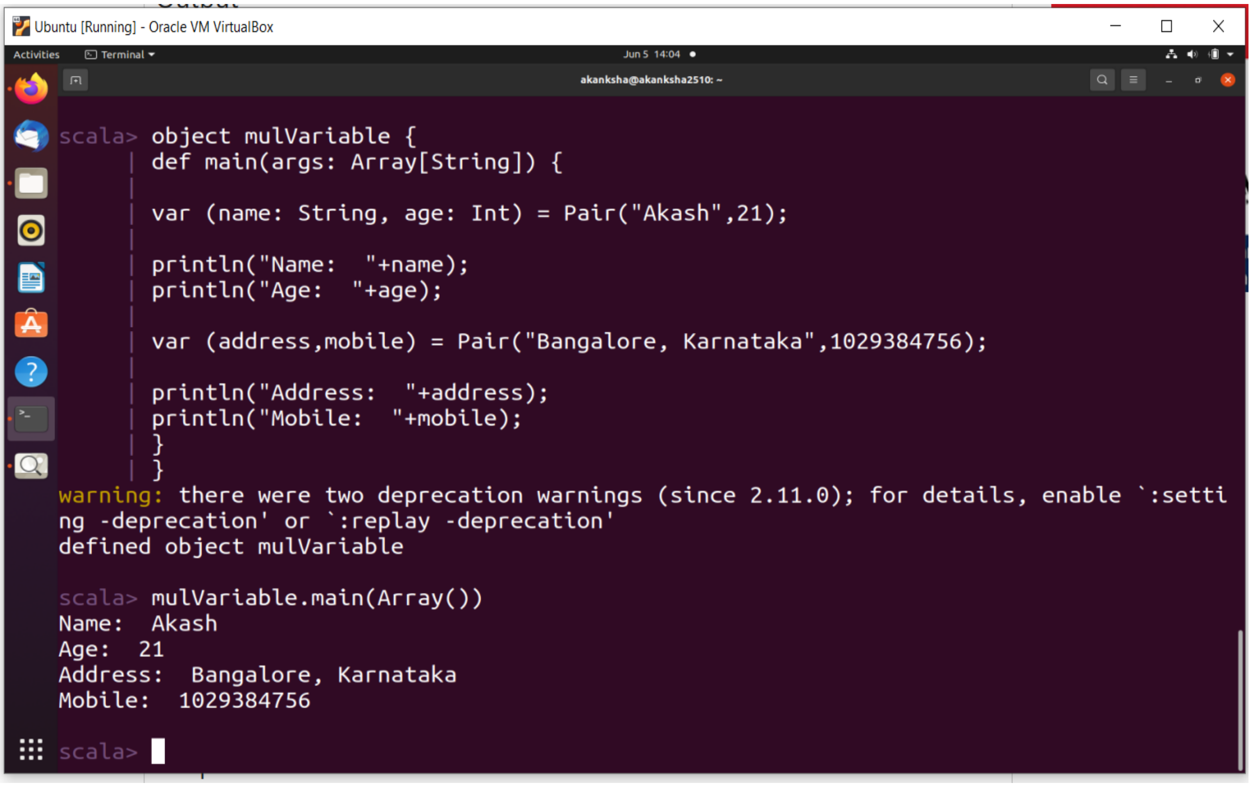
```
scala> findLargest.main(Array())
Largest number is :  65

scala> object checkNumber {
     | def main(args: Array[String]) {
     |
     | var number = (-2021);
     |
     | if(number==0){
     | println("Number is zero");
     | }
     | else if(number>0) {
     | println("Number is positive");
     | }
     | else {
     | println("Number is negative");
     | }
     | }
     | }
defined object checkNumber

scala> checkNumber.main(Array())
Number is negative

scala>
```

Fig 9.3



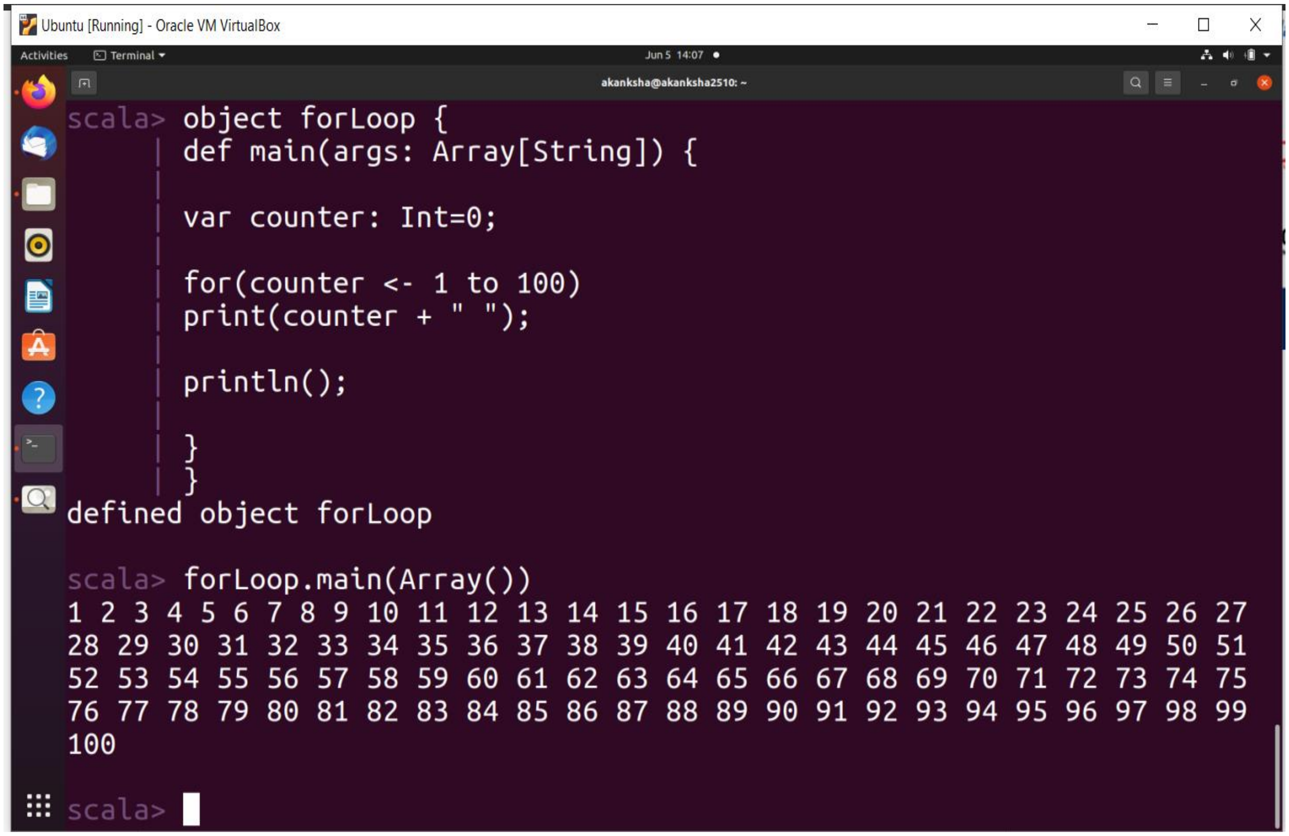
```
     | println("Number is positive");
     | }
     | else {
     | println("Number is negative");
     | }
     | }
     | }
defined object checkNumber

scala> checkNumber.main(Array())
Number is negative

scala> object printString {
     | def main(args: Array[String]) {
     |
     | val text : String = "It is a spark shell.";
     |
     | println("String is :  "+text);
     | }
     | }
defined object printString

scala> printString.main(Array())
String is :  It is a spark shell.

scala> █
```

Fig 9.4



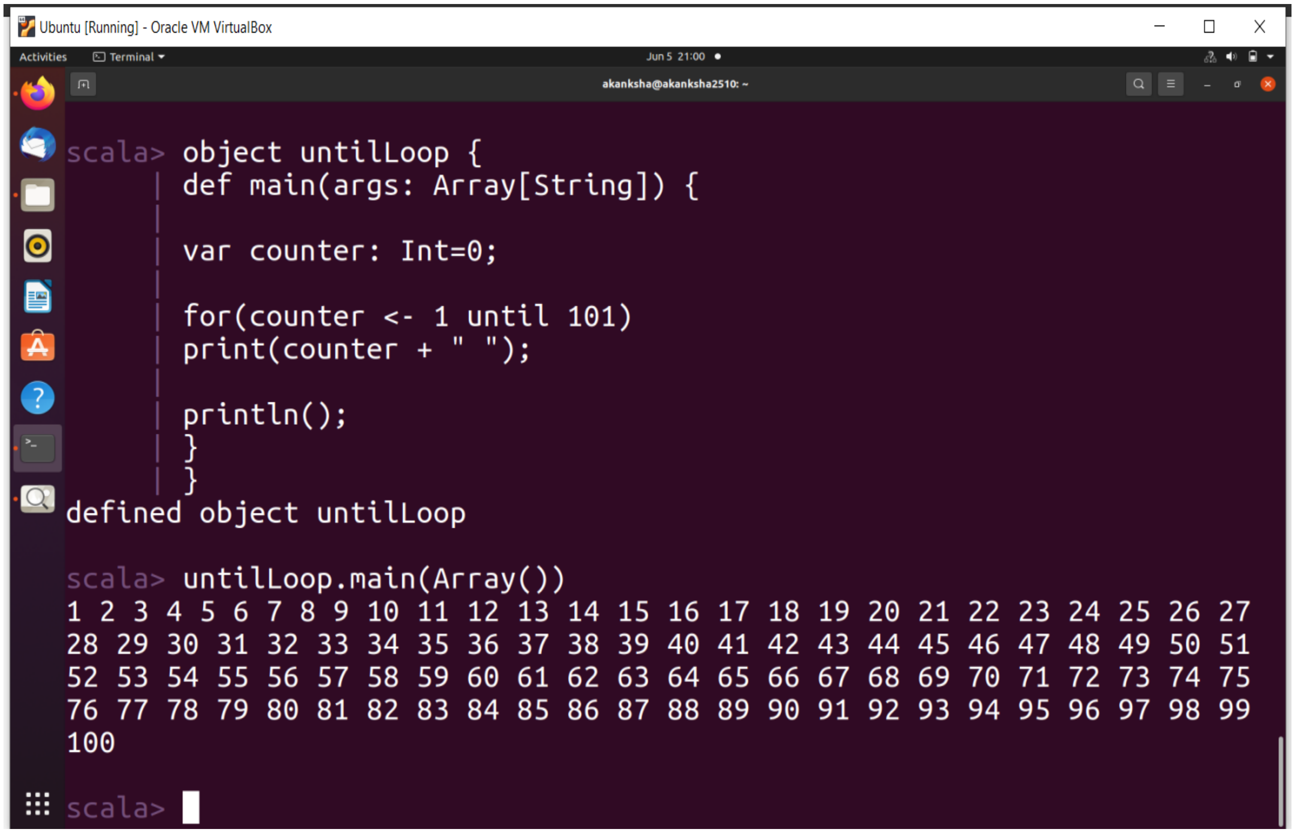
```
scala> object mulVariable {
     | def main(args: Array[String]) {
     |
     | var (name: String, age: Int) = Pair("Akash",21);
     |
     | println("Name:  "+name);
     | println("Age:  "+age);
     |
     | var (address,mobile) = Pair("Bangalore, Karnataka",1029384756);
     |
     | println("Address:  "+address);
     | println("Mobile:  "+mobile);
     | }
     | }
warning: there were two deprecation warnings (since 2.11.0); for details, enable `:setti
ng -deprecation' or `:replay -deprecation'
defined object mulVariable

scala> mulVariable.main(Array())
Name:  Akash
Age:  21
Address:  Bangalore, Karnataka
Mobile:  1029384756

scala> 
```

Fig 9.5



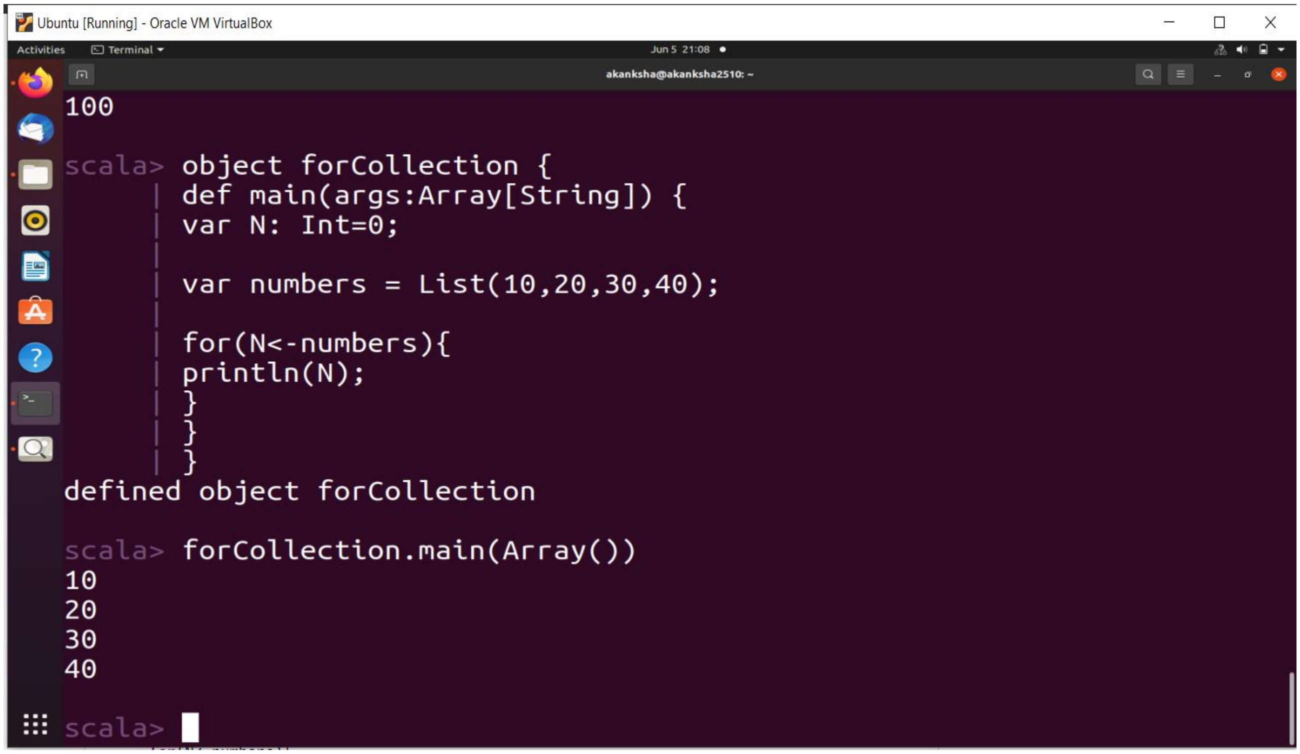
```
scala> object forLoop {
     | def main(args: Array[String]) {
     |
     | var counter: Int=0;
     |
     | for(counter <- 1 to 100)
     | print(counter + " ");
     |
     | println();
     |
     | }
     | }
defined object forLoop

scala> forLoop.main(Array())
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
100

scala> 
```

Fig 9.6



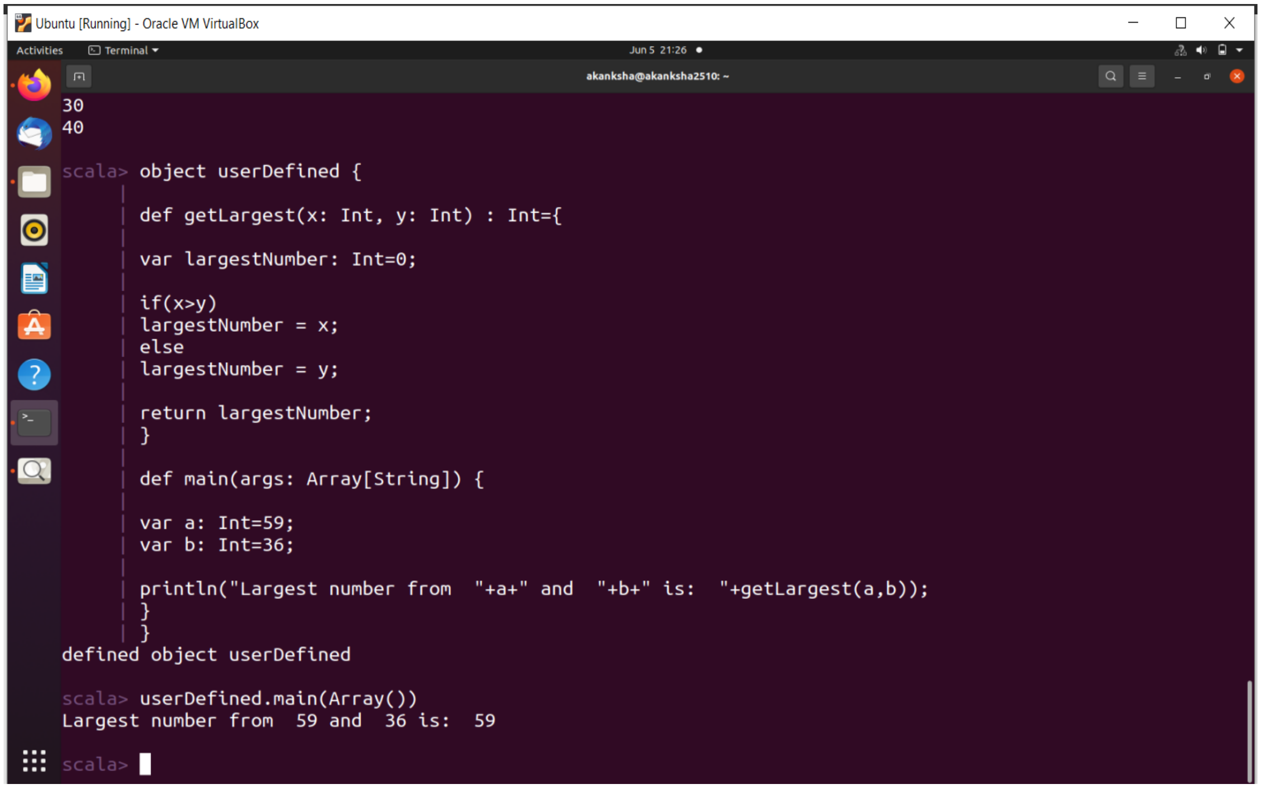
```
scala> object untilLoop {
     | def main(args: Array[String]) {
     |
     | var counter: Int=0;
     |
     | for(counter <- 1 until 101)
     | print(counter + " ");
     |
     | println();
     | }
     | }
defined object untilLoop

scala> untilLoop.main(Array())
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51
52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99
100

scala> 
```

Fig 9.7



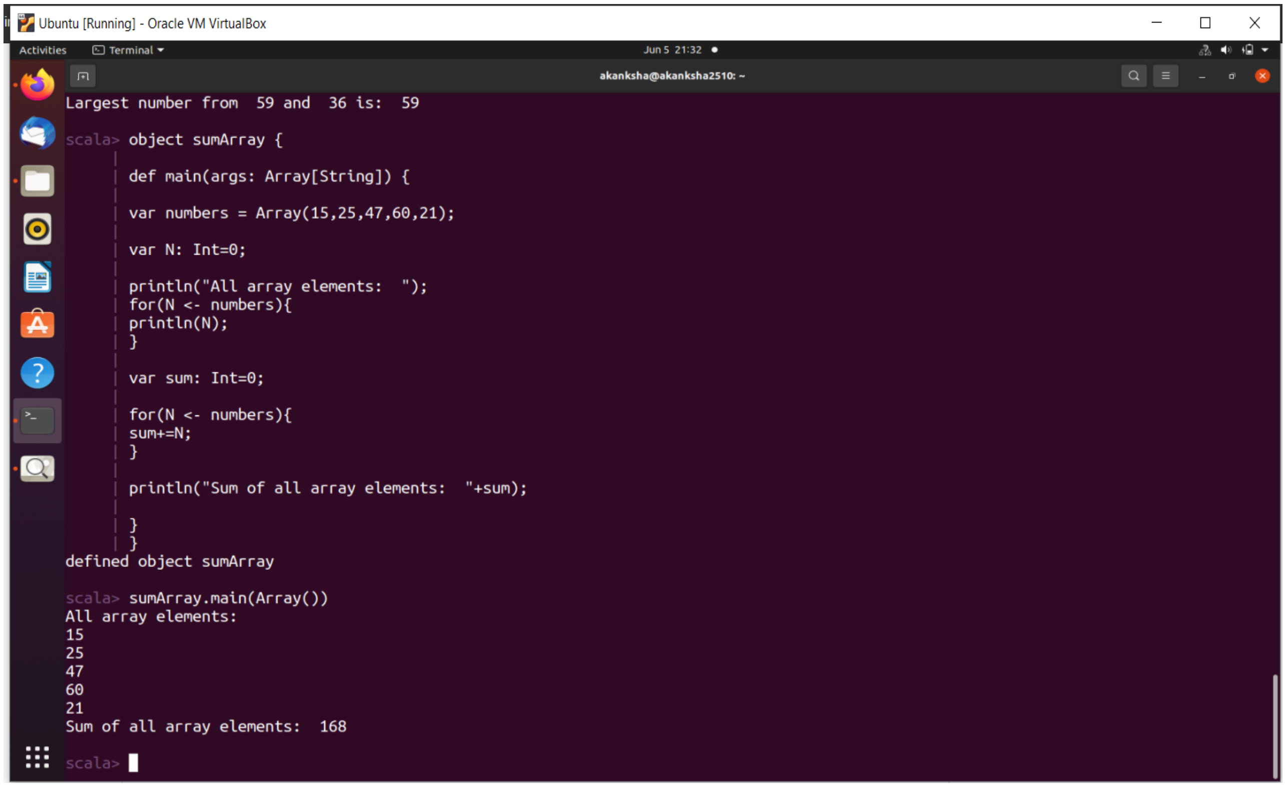
```
100

scala> object forCollection {
     | def main(args:Array[String]) {
     | var N: Int=0;
     |
     | var numbers = List(10,20,30,40);
     |
     | for(N<-numbers){
     | println(N);
     | }
     | }
     | }
defined object forCollection

scala> forCollection.main(Array())
10
20
30
40

scala> 
```

Fig 9.8



```
30
40

scala> object userDefined {
     |
     | def getLargest(x: Int, y: Int) : Int={
     |
     | var largestNumber: Int=0;
     |
     | if(x>y)
     | largestNumber = x;
     | else
     | largestNumber = y;
     |
     | return largestNumber;
     | }
     |
     | def main(args: Array[String]) {
     |
     | var a: Int=59;
     | var b: Int=36;
     |
     | println("Largest number from  "+a+" and  "+b+" is:  "+getLargest(a,b));
     | }
     | }
defined object userDefined

scala> userDefined.main(Array())
Largest number from  59 and  36 is:  59

scala> 
```

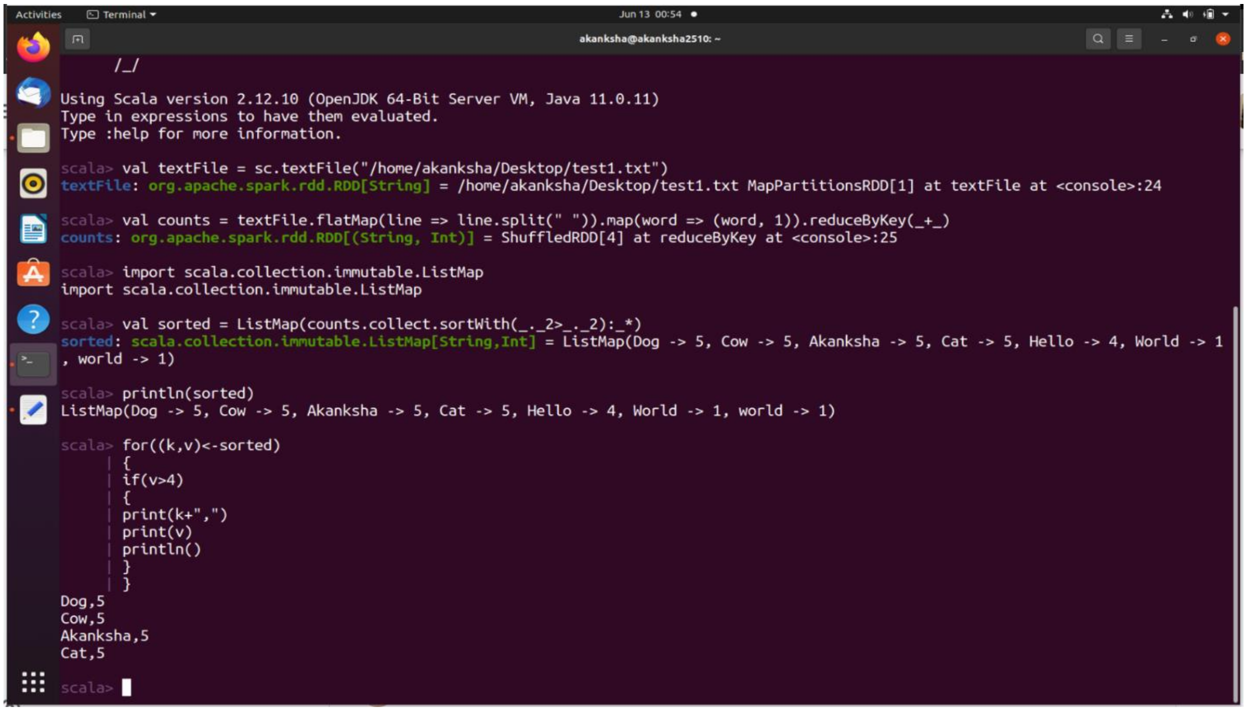
Fig 9.9



```
Largest number from  59 and  36 is:  59

scala> object sumArray {
     |
     | def main(args: Array[String]) {
     |
     | var numbers = Array(15,25,47,60,21);
     |
     | var N: Int=0;
     |
     | println("All array elements:  ");
     | for(N <- numbers){
     | println(N);
     | }
     |
     | var sum: Int=0;
     |
     | for(N <- numbers){
     | sum+=N;
     | }
     |
     | println("Sum of all array elements:  "+sum);
     |
     | }
     | }
defined object sumArray

scala> sumArray.main(Array())
All array elements:
15
25
47
60
21
Sum of all array elements:  168

scala> 
```

Fig 9.10

Program 10. Using RDD and Flat Map count how many times each word appears in a file and write out a list of words whose count is strictly greater than 4 using Spark.



```
      /_/

Using Scala version 2.12.10 (OpenJDK 64-Bit Server VM, Java 11.0.11)
Type in expressions to have them evaluated.
Type :help for more information.

scala> val textFile = sc.textFile("/home/akanksha/Desktop/test1.txt")
textFile: org.apache.spark.rdd.RDD[String] = /home/akanksha/Desktop/test1.txt MapPartitionsRDD[1] at textFile at <console>:24

scala> val counts = textFile.flatMap(line => line.split(" ")).map(word => (word, 1)).reduceByKey(_+_)
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[4] at reduceByKey at <console>:25

scala> import scala.collection.immutable.ListMap
import scala.collection.immutable.ListMap

scala> val sorted = ListMap(counts.collect.sortWith(_._2>_._2):_*)
sorted: scala.collection.immutable.ListMap[String,Int] = ListMap(Dog -> 5, Cow -> 5, Akanksha -> 5, Cat -> 5, Hello -> 4, World -> 1
, world -> 1)

scala> println(sorted)
ListMap(Dog -> 5, Cow -> 5, Akanksha -> 5, Cat -> 5, Hello -> 4, World -> 1, world -> 1)

scala> for((k,v)<-sorted)
     | {
     | if(v>4)
     | {
     | print(k+",")
     | print(v)
     | println()
     | }
     | }
Dog,5
Cow,5
Akanksha,5
Cat,5

scala> 
```

Fig 10.1