VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"JnanaSangama", Belgaum -590014, Karnataka.

LAB REPORT
on

BIG DATA ANALYTICS
(20CS6PEBDA)

Submitted by

DISHA N (1BM19CS051)

in partial fulfillment for the award of the degree of
BACHELOR OF ENGINEERING
in
COMPUTER SCIENCE AND ENGINEERING

B.M.S. COLLEGE OF ENGINEERING
(Autonomous Institution under VTU)
BENGALURU-560019
May-2022 to July-2022

B. M. S. College of Engineering,
Bull Temple Road, Bangalore 560019
(Affiliated To Visvesvaraya Technological University, Belgaum)
Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the Lab work entitled "**BIG DATA ANALYTICS**" carried out by **DISHA N (1BM19CS051), who is a bonafide student of B. M. S. College of Engineering.** It is in partial fulfillment for the award of **Bachelor of Engineering in Computer Science and Engineering** of the Visvesvaraya Technological University, Belgaum during the year 2022. The Lab report has been approved as it satisfies the academic requirements in respect of a **BIG DATA ANALYTICS - (20CS6PEBDA)**work prescribed for the said degree.

Dr. PALLAVI G B
Assistant Professor
Department of CSE
BMSCE, Bengaluru

Dr. Jyothi S Nayak
Professor and Head
Department of CSE
BMSCE, Bengaluru

Index Sheet

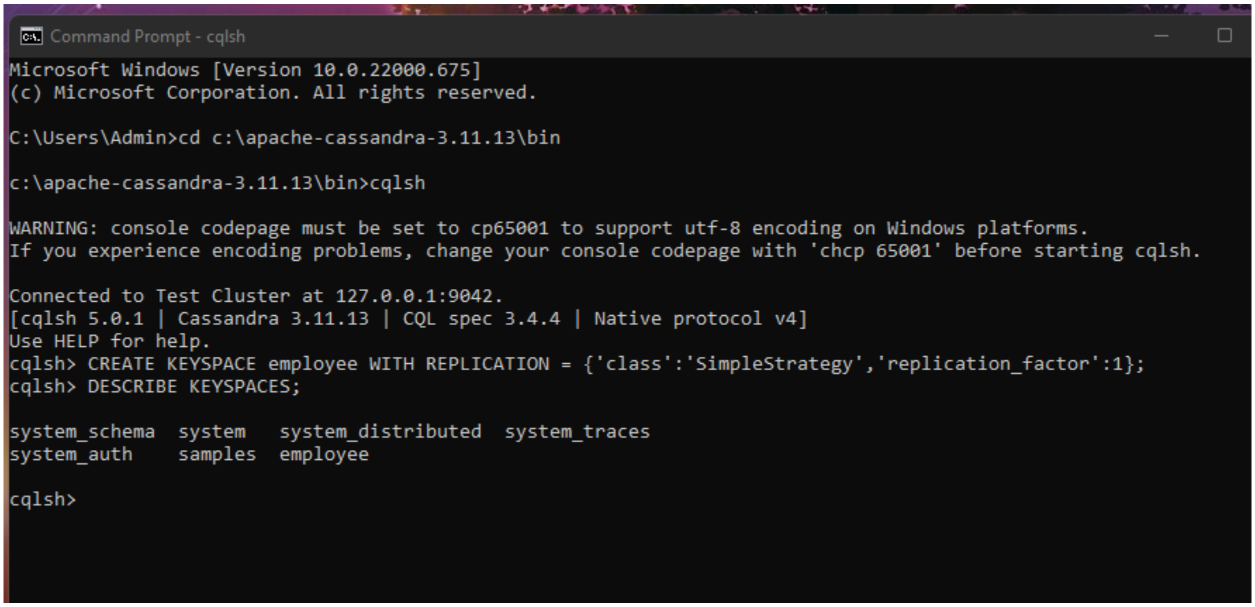
Course Outcome

CO1	Apply the concept of NoSQL, Hadoop or Spark for a given task
CO2	Analyze the Big Data and obtain insight using data analytics mechanisms.
CO3	Design and implement Big data applications by applying NoSQL, Hadoop or Spark

Cassandra Lab Program 1: -

Perform the following DB operations using Cassandra.

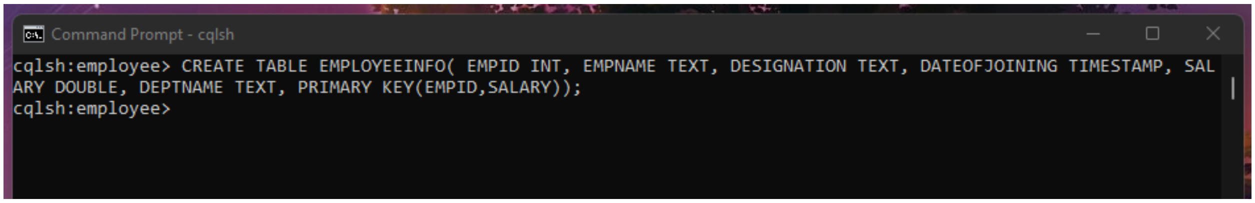
1. Create a key space by name Employee



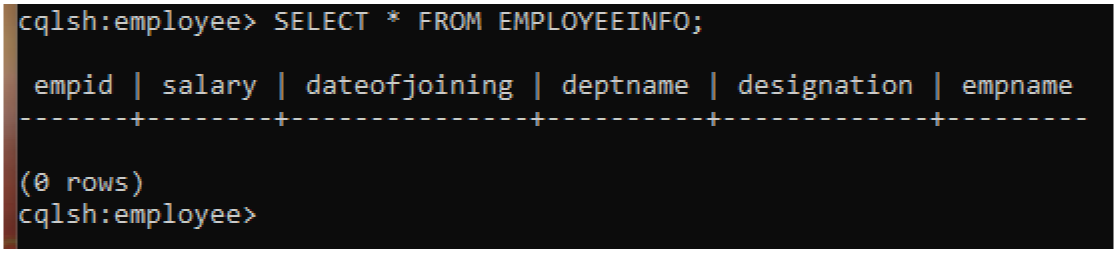
```
Microsoft Windows [Version 10.0.22000.675]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Admin>cd c:\apache-cassandra-3.11.13\bin

c:\apache-cassandra-3.11.13\bin>cqlsh

WARNING: console codepage must be set to cp65001 to support utf-8 encoding on Windows platforms.
If you experience encoding problems, change your console codepage with 'chcp 65001' before starting cqlsh.

Connected to Test Cluster at 127.0.0.1:9042.
[cqlsh 5.0.1 | Cassandra 3.11.13 | CQL spec 3.4.4 | Native protocol v4]
Use HELP for help.
cqlsh> CREATE KEYSPACE employee WITH REPLICATION = {'class':'SimpleStrategy','replication_factor':1};
cqlsh> DESCRIBE KEYSPACES;

system_schema  system   system_distributed  system_traces
system_auth    samples  employee

cqlsh>
```

2. Create a column family by name Employee-Info with attributes Emp_Id Primary Key, Emp_Name,

Designation, Date_of_Joining, Salary, Dept_Name

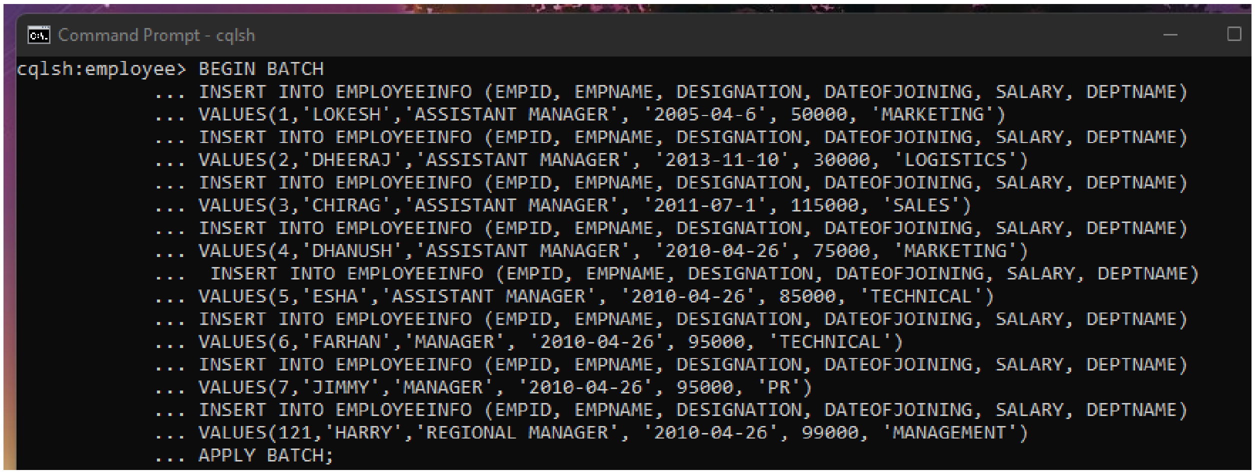


```
cqlsh:employee> CREATE TABLE EMPLOYEEINFO( EMPID INT, EMPNAME TEXT, DESIGNATION TEXT, DATEOFJOINING TIMESTAMP, SAL
ARY DOUBLE, DEPTNAME TEXT, PRIMARY KEY(EMPID,SALARY));
cqlsh:employee>
```

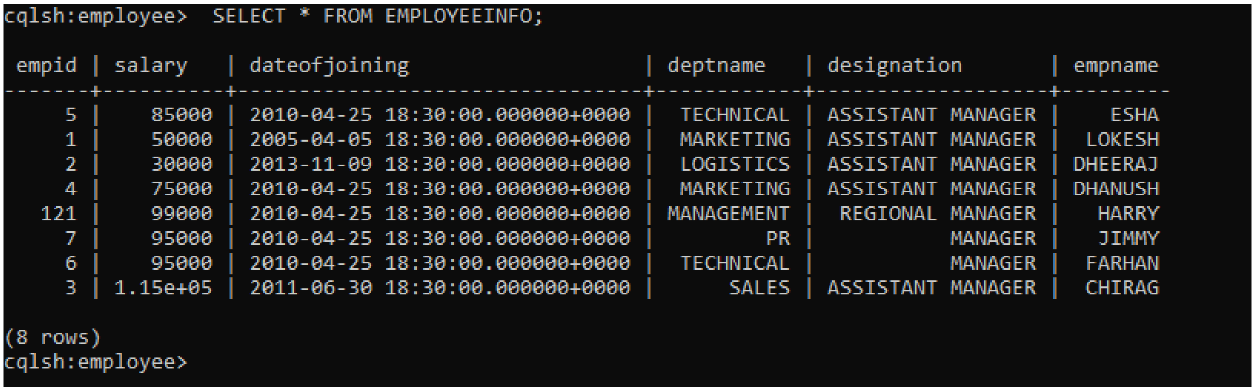


```
cqlsh:employee> SELECT * FROM EMPLOYEEINFO;

 empid | salary | dateofjoining | deptname | designation | empname
-------+--------+---------------+----------+-------------+---------

(0 rows)
cqlsh:employee>
```

3. Insert the values into the table in batch

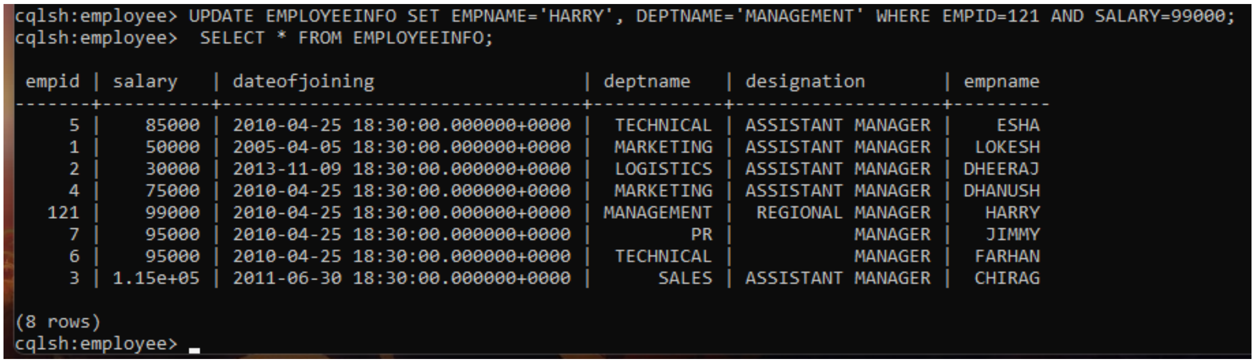


```
cqlsh:employee> BEGIN BATCH
            ... INSERT INTO EMPLOYEEINFO (EMPID, EMPNAME, DESIGNATION, DATEOFJOINING, SALARY, DEPTNAME)
            ... VALUES(1,'LOKESH','ASSISTANT MANAGER', '2005-04-6', 50000, 'MARKETING')
            ... INSERT INTO EMPLOYEEINFO (EMPID, EMPNAME, DESIGNATION, DATEOFJOINING, SALARY, DEPTNAME)
            ... VALUES(2,'DHEERAJ','ASSISTANT MANAGER', '2013-11-10', 30000, 'LOGISTICS')
            ... INSERT INTO EMPLOYEEINFO (EMPID, EMPNAME, DESIGNATION, DATEOFJOINING, SALARY, DEPTNAME)
            ... VALUES(3,'CHIRAG','ASSISTANT MANAGER', '2011-07-1', 115000, 'SALES')
            ... INSERT INTO EMPLOYEEINFO (EMPID, EMPNAME, DESIGNATION, DATEOFJOINING, SALARY, DEPTNAME)
            ... VALUES(4,'DHANUSH','ASSISTANT MANAGER', '2010-04-26', 75000, 'MARKETING')
            ...  INSERT INTO EMPLOYEEINFO (EMPID, EMPNAME, DESIGNATION, DATEOFJOINING, SALARY, DEPTNAME)
            ... VALUES(5,'ESHA','ASSISTANT MANAGER', '2010-04-26', 85000, 'TECHNICAL')
            ... INSERT INTO EMPLOYEEINFO (EMPID, EMPNAME, DESIGNATION, DATEOFJOINING, SALARY, DEPTNAME)
            ... VALUES(6,'FARHAN','MANAGER', '2010-04-26', 95000, 'TECHNICAL')
            ... INSERT INTO EMPLOYEEINFO (EMPID, EMPNAME, DESIGNATION, DATEOFJOINING, SALARY, DEPTNAME)
            ... VALUES(7,'JIMMY','MANAGER', '2010-04-26', 95000, 'PR')
            ... INSERT INTO EMPLOYEEINFO (EMPID, EMPNAME, DESIGNATION, DATEOFJOINING, SALARY, DEPTNAME)
            ... VALUES(121,'HARRY','REGIONAL MANAGER', '2010-04-26', 99000, 'MANAGEMENT')
            ... APPLY BATCH;
```



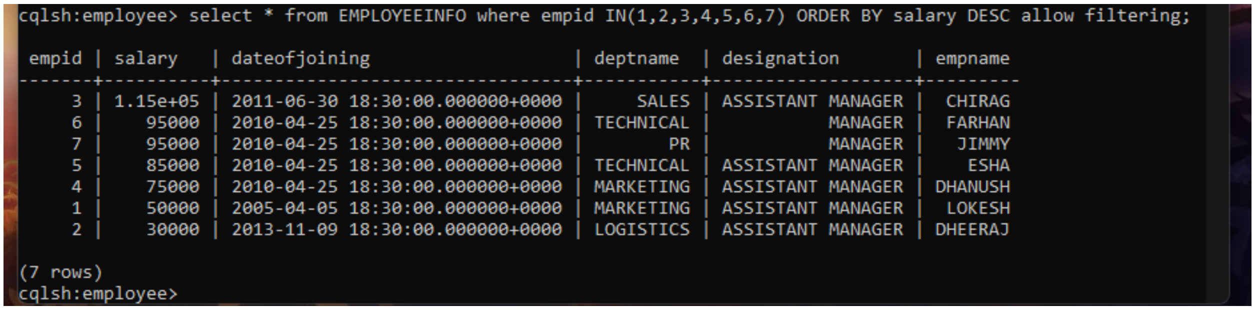
```
cqlsh:employee>  SELECT * FROM EMPLOYEEINFO;

 empid | salary   | dateofjoining                   | deptname   | designation       | empname
-------+----------+---------------------------------+------------+-------------------+---------
     5 |    85000 | 2010-04-25 18:30:00.000000+0000 |  TECHNICAL | ASSISTANT MANAGER |    ESHA
     1 |    50000 | 2005-04-05 18:30:00.000000+0000 |  MARKETING | ASSISTANT MANAGER |  LOKESH
     2 |    30000 | 2013-11-09 18:30:00.000000+0000 |  LOGISTICS | ASSISTANT MANAGER | DHEERAJ
     4 |    75000 | 2010-04-25 18:30:00.000000+0000 |  MARKETING | ASSISTANT MANAGER | DHANUSH
   121 |    99000 | 2010-04-25 18:30:00.000000+0000 | MANAGEMENT |  REGIONAL MANAGER |   HARRY
     7 |    95000 | 2010-04-25 18:30:00.000000+0000 |         PR |           MANAGER |   JIMMY
     6 |    95000 | 2010-04-25 18:30:00.000000+0000 |  TECHNICAL |           MANAGER |  FARHAN
     3 | 1.15e+05 | 2011-06-30 18:30:00.000000+0000 |      SALES | ASSISTANT MANAGER |  CHIRAG

(8 rows)
cqlsh:employee>
```

4. Update Employee name and Department of Emp-Id 121



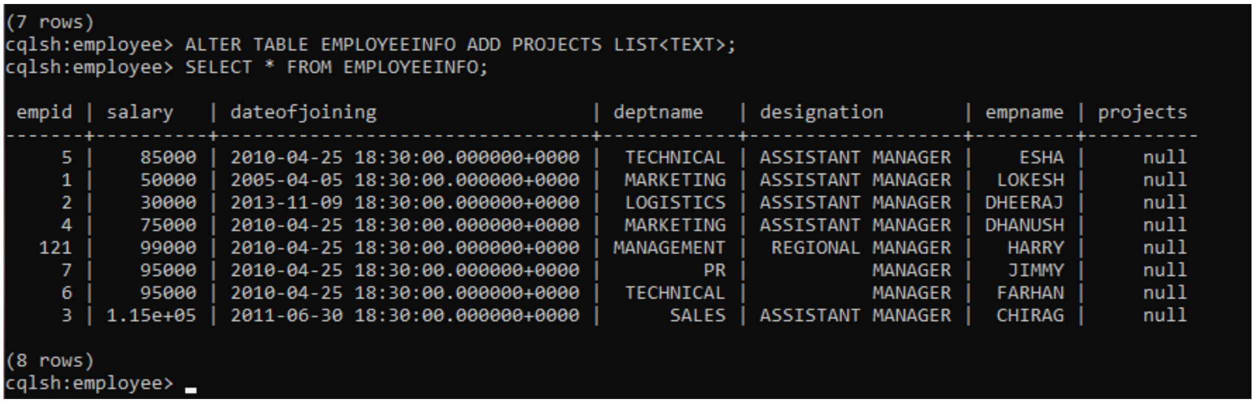
```
cqlsh:employee> UPDATE EMPLOYEEINFO SET EMPNAME='HARRY', DEPTNAME='MANAGEMENT' WHERE EMPID=121 AND SALARY=99000;
cqlsh:employee>  SELECT * FROM EMPLOYEEINFO;

 empid | salary   | dateofjoining                   | deptname   | designation       | empname
-------+----------+---------------------------------+------------+-------------------+---------
     5 |    85000 | 2010-04-25 18:30:00.000000+0000 |  TECHNICAL | ASSISTANT MANAGER |    ESHA
     1 |    50000 | 2005-04-05 18:30:00.000000+0000 |  MARKETING | ASSISTANT MANAGER |  LOKESH
     2 |    30000 | 2013-11-09 18:30:00.000000+0000 |  LOGISTICS | ASSISTANT MANAGER | DHEERAJ
     4 |    75000 | 2010-04-25 18:30:00.000000+0000 |  MARKETING | ASSISTANT MANAGER | DHANUSH
   121 |    99000 | 2010-04-25 18:30:00.000000+0000 | MANAGEMENT |  REGIONAL MANAGER |   HARRY
     7 |    95000 | 2010-04-25 18:30:00.000000+0000 |         PR |           MANAGER |   JIMMY
     6 |    95000 | 2010-04-25 18:30:00.000000+0000 |  TECHNICAL |           MANAGER |  FARHAN
     3 | 1.15e+05 | 2011-06-30 18:30:00.000000+0000 |      SALES | ASSISTANT MANAGER |  CHIRAG

(8 rows)
cqlsh:employee> _
```

5. Sort the details of Employee records based on salary (Note:- cql>PAGING OFF)



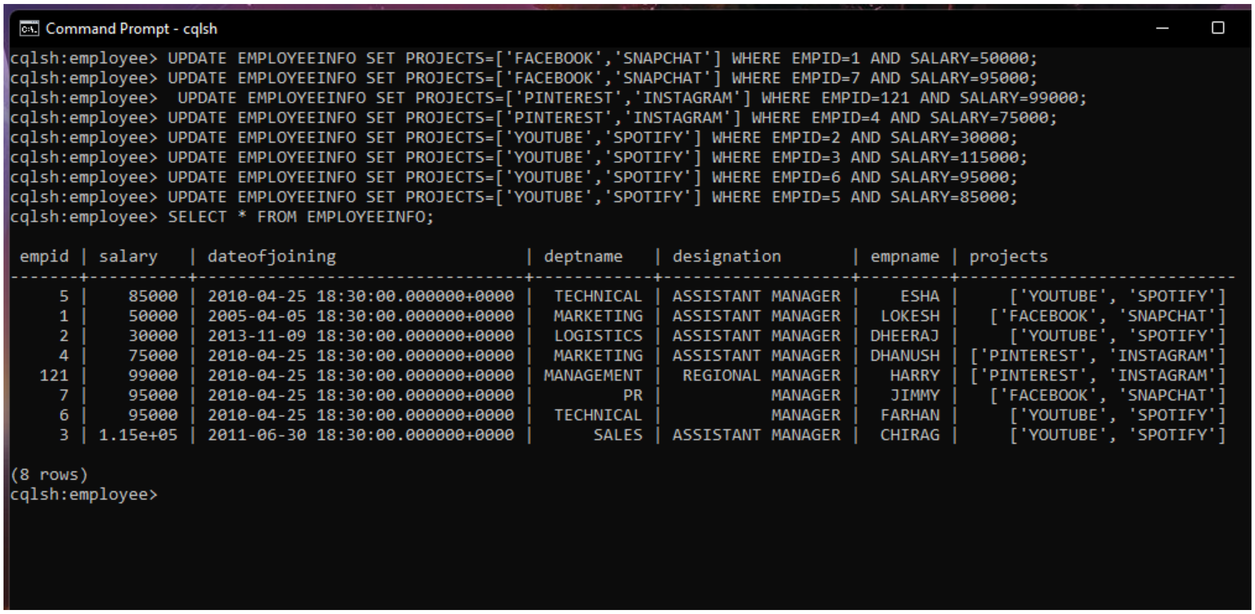
```
cqlsh:employee> select * from EMPLOYEEINFO where empid IN(1,2,3,4,5,6,7) ORDER BY salary DESC allow filtering;

 empid | salary   | dateofjoining                   | deptname  | designation       | empname
-------+----------+---------------------------------+-----------+-------------------+---------
     3 | 1.15e+05 | 2011-06-30 18:30:00.000000+0000 |     SALES | ASSISTANT MANAGER |  CHIRAG
     6 |    95000 | 2010-04-25 18:30:00.000000+0000 | TECHNICAL |           MANAGER |  FARHAN
     7 |    95000 | 2010-04-25 18:30:00.000000+0000 |        PR |           MANAGER |   JIMMY
     5 |    85000 | 2010-04-25 18:30:00.000000+0000 | TECHNICAL | ASSISTANT MANAGER |    ESHA
     4 |    75000 | 2010-04-25 18:30:00.000000+0000 | MARKETING | ASSISTANT MANAGER | DHANUSH
     1 |    50000 | 2005-04-05 18:30:00.000000+0000 | MARKETING | ASSISTANT MANAGER |  LOKESH
     2 |    30000 | 2013-11-09 18:30:00.000000+0000 | LOGISTICS | ASSISTANT MANAGER | DHEERAJ

(7 rows)
cqlsh:employee>
```

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 rows)
cqlsh:employee> ALTER TABLE EMPLOYEEINFO ADD PROJECTS LIST<TEXT>;
cqlsh:employee> SELECT * FROM EMPLOYEEINFO;

 empid | salary   | dateofjoining                   | deptname   | designation       | empname | projects
-------+----------+---------------------------------+------------+-------------------+---------+----------
     5 |    85000 | 2010-04-25 18:30:00.000000+0000 |  TECHNICAL | ASSISTANT MANAGER |    ESHA |     null
     1 |    50000 | 2005-04-05 18:30:00.000000+0000 |  MARKETING | ASSISTANT MANAGER |  LOKESH |     null
     2 |    30000 | 2013-11-09 18:30:00.000000+0000 |  LOGISTICS | ASSISTANT MANAGER | DHEERAJ |     null
     4 |    75000 | 2010-04-25 18:30:00.000000+0000 |  MARKETING | ASSISTANT MANAGER | DHANUSH |     null
   121 |    99000 | 2010-04-25 18:30:00.000000+0000 | MANAGEMENT |  REGIONAL MANAGER |   HARRY |     null
     7 |    95000 | 2010-04-25 18:30:00.000000+0000 |         PR |           MANAGER |   JIMMY |     null
     6 |    95000 | 2010-04-25 18:30:00.000000+0000 |  TECHNICAL |           MANAGER |  FARHAN |     null
     3 | 1.15e+05 | 2011-06-30 18:30:00.000000+0000 |      SALES | ASSISTANT MANAGER |  CHIRAG |     null

(8 rows)
cqlsh:employee>
```

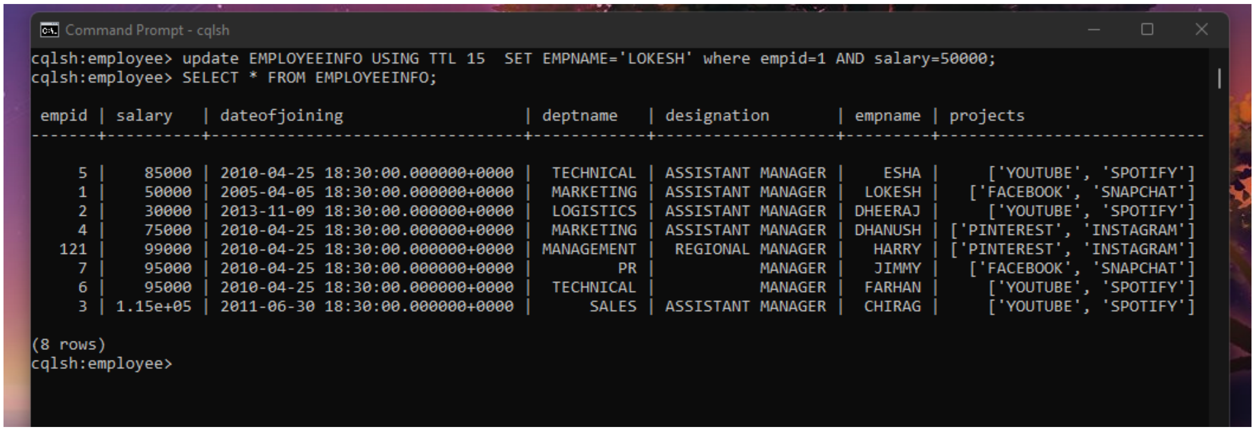
7. Update the altered table to add project names.



```
cqlsh:employee> UPDATE EMPLOYEEINFO SET PROJECTS=['FACEBOOK','SNAPCHAT'] WHERE EMPID=1 AND SALARY=50000;
cqlsh:employee> UPDATE EMPLOYEEINFO SET PROJECTS=['FACEBOOK','SNAPCHAT'] WHERE EMPID=7 AND SALARY=95000;
cqlsh:employee>  UPDATE EMPLOYEEINFO SET PROJECTS=['PINTEREST','INSTAGRAM'] WHERE EMPID=121 AND SALARY=99000;
cqlsh:employee> UPDATE EMPLOYEEINFO SET PROJECTS=['PINTEREST','INSTAGRAM'] WHERE EMPID=4 AND SALARY=75000;
cqlsh:employee> UPDATE EMPLOYEEINFO SET PROJECTS=['YOUTUBE','SPOTIFY'] WHERE EMPID=2 AND SALARY=30000;
cqlsh:employee> UPDATE EMPLOYEEINFO SET PROJECTS=['YOUTUBE','SPOTIFY'] WHERE EMPID=3 AND SALARY=115000;
cqlsh:employee> UPDATE EMPLOYEEINFO SET PROJECTS=['YOUTUBE','SPOTIFY'] WHERE EMPID=6 AND SALARY=95000;
cqlsh:employee> UPDATE EMPLOYEEINFO SET PROJECTS=['YOUTUBE','SPOTIFY'] WHERE EMPID=5 AND SALARY=85000;
cqlsh:employee> SELECT * FROM EMPLOYEEINFO;

 empid | salary   | dateofjoining                   | deptname   | designation       | empname | projects
-------+----------+---------------------------------+------------+-------------------+---------+----------------------------
     5 |    85000 | 2010-04-25 18:30:00.000000+0000 |  TECHNICAL | ASSISTANT MANAGER |    ESHA |      ['YOUTUBE', 'SPOTIFY']
     1 |    50000 | 2005-04-05 18:30:00.000000+0000 |  MARKETING | ASSISTANT MANAGER |  LOKESH |    ['FACEBOOK', 'SNAPCHAT']
     2 |    30000 | 2013-11-09 18:30:00.000000+0000 |  LOGISTICS | ASSISTANT MANAGER | DHEERAJ |      ['YOUTUBE', 'SPOTIFY']
     4 |    75000 | 2010-04-25 18:30:00.000000+0000 |  MARKETING | ASSISTANT MANAGER | DHANUSH | ['PINTEREST', 'INSTAGRAM']
   121 |    99000 | 2010-04-25 18:30:00.000000+0000 | MANAGEMENT |  REGIONAL MANAGER |   HARRY | ['PINTEREST', 'INSTAGRAM']
     7 |    95000 | 2010-04-25 18:30:00.000000+0000 |         PR |           MANAGER |   JIMMY |    ['FACEBOOK', 'SNAPCHAT']
     6 |    95000 | 2010-04-25 18:30:00.000000+0000 |  TECHNICAL |           MANAGER |  FARHAN |      ['YOUTUBE', 'SPOTIFY']
     3 | 1.15e+05 | 2011-06-30 18:30:00.000000+0000 |      SALES | ASSISTANT MANAGER |  CHIRAG |      ['YOUTUBE', 'SPOTIFY']

(8 rows)
cqlsh:employee>
```

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

//BEFORE 15 seconds

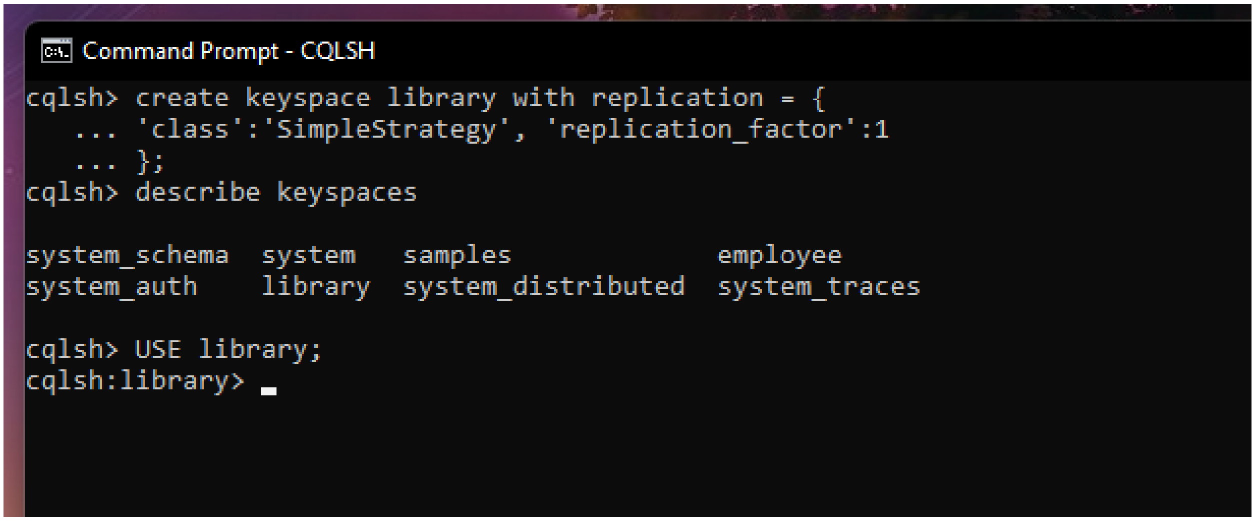


```
cqlsh:employee> update EMPLOYEEINFO USING TTL 15  SET EMPNAME='LOKESH' where empid=1 AND salary=50000;
cqlsh:employee> SELECT * FROM EMPLOYEEINFO;

 empid | salary   | dateofjoining                   | deptname   | designation       | empname | projects
-------+----------+---------------------------------+------------+-------------------+---------+----------------------------

     5 |    85000 | 2010-04-25 18:30:00.000000+0000 |  TECHNICAL | ASSISTANT MANAGER |    ESHA |      ['YOUTUBE', 'SPOTIFY']
     1 |    50000 | 2005-04-05 18:30:00.000000+0000 |  MARKETING | ASSISTANT MANAGER |  LOKESH |    ['FACEBOOK', 'SNAPCHAT']
     2 |    30000 | 2013-11-09 18:30:00.000000+0000 |  LOGISTICS | ASSISTANT MANAGER | DHEERAJ |      ['YOUTUBE', 'SPOTIFY']
     4 |    75000 | 2010-04-25 18:30:00.000000+0000 |  MARKETING | ASSISTANT MANAGER | DHANUSH | ['PINTEREST', 'INSTAGRAM']
   121 |    99000 | 2010-04-25 18:30:00.000000+0000 | MANAGEMENT |  REGIONAL MANAGER |   HARRY | ['PINTEREST', 'INSTAGRAM']
     7 |    95000 | 2010-04-25 18:30:00.000000+0000 |         PR |           MANAGER |   JIMMY |    ['FACEBOOK', 'SNAPCHAT']
     6 |    95000 | 2010-04-25 18:30:00.000000+0000 |  TECHNICAL |           MANAGER |  FARHAN |      ['YOUTUBE', 'SPOTIFY']
     3 | 1.15e+05 | 2011-06-30 18:30:00.000000+0000 |      SALES | ASSISTANT MANAGER |  CHIRAG |      ['YOUTUBE', 'SPOTIFY']

(8 rows)
cqlsh:employee>
```

Cassandra Lab Program 2: -

Perform the following DB operations using Cassandra.

1.Create a key space by name Library

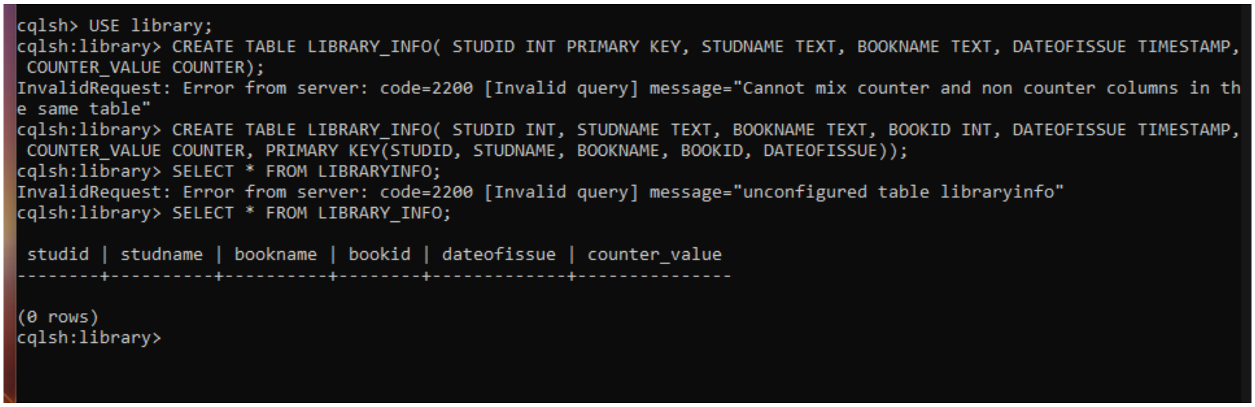


```
cqlsh> create keyspace library with replication = {
   ... 'class':'SimpleStrategy', 'replication_factor':1
   ... };
cqlsh> describe keyspaces

system_schema  system   samples             employee
system_auth    library  system_distributed  system_traces

cqlsh> USE library;
cqlsh:library> _
```

2. Create a column family by name Library-Info with attributes Stud_Id Primary Key,

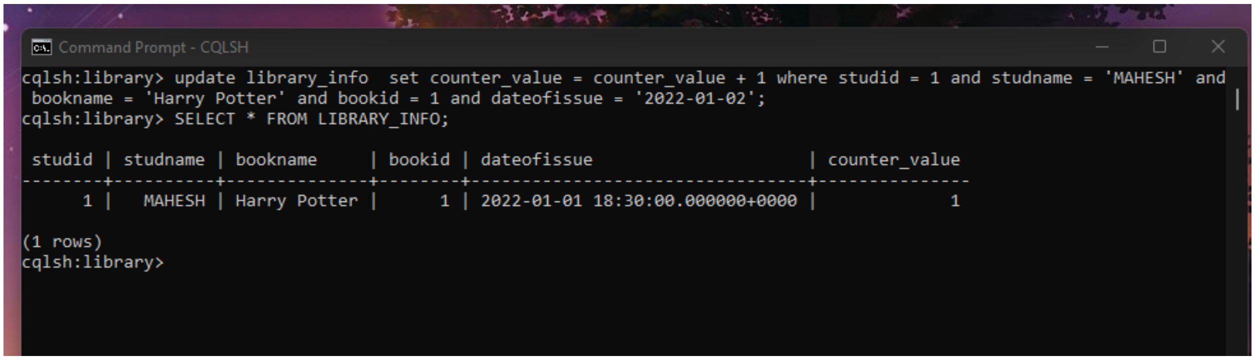
Counter_value of type Counter,

Stud_Name, Book-Name, Book-Id, Date_of_issue

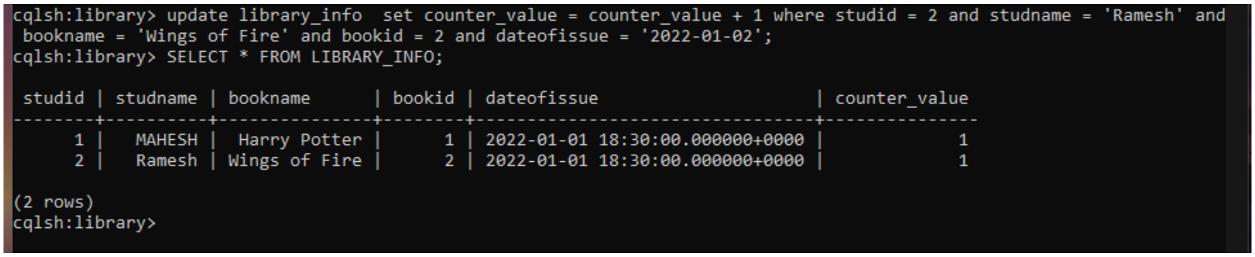


```
cqlsh> USE library;
cqlsh:library> CREATE TABLE LIBRARY_INFO( STUDID INT PRIMARY KEY, STUDNAME TEXT, BOOKNAME TEXT, DATEOFISSUE TIMESTAMP,
 COUNTER_VALUE COUNTER);
InvalidRequest: Error from server: code=2200 [Invalid query] message="Cannot mix counter and non counter columns in th
e same table"
cqlsh:library> CREATE TABLE LIBRARY_INFO( STUDID INT, STUDNAME TEXT, BOOKNAME TEXT, BOOKID INT, DATEOFISSUE TIMESTAMP,
 COUNTER_VALUE COUNTER, PRIMARY KEY(STUDID, STUDNAME, BOOKNAME, BOOKID, DATEOFISSUE));
cqlsh:library> SELECT * FROM LIBRARYINFO;
InvalidRequest: Error from server: code=2200 [Invalid query] message="unconfigured table libraryinfo"
cqlsh:library> SELECT * FROM LIBRARY_INFO;

 studid | studname | bookname | bookid | dateofissue | counter_value
--------+----------+----------+--------+-------------+---------------

(0 rows)
cqlsh:library>
```

3.Insert the values into the table in batch

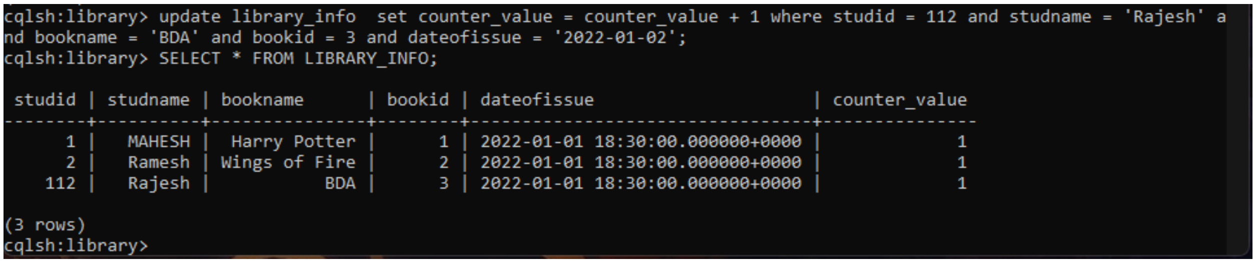


```
cqlsh:library> update library_info  set counter_value = counter_value + 1 where studid = 1 and studname = 'MAHESH' and
 bookname = 'Harry Potter' and bookid = 1 and dateofissue = '2022-01-02';
cqlsh:library> SELECT * FROM LIBRARY_INFO;

 studid | studname | bookname     | bookid | dateofissue                     | counter_value
--------+----------+--------------+--------+---------------------------------+---------------
      1 |   MAHESH | Harry Potter |      1 | 2022-01-01 18:30:00.000000+0000 |             1

(1 rows)
cqlsh:library>
```

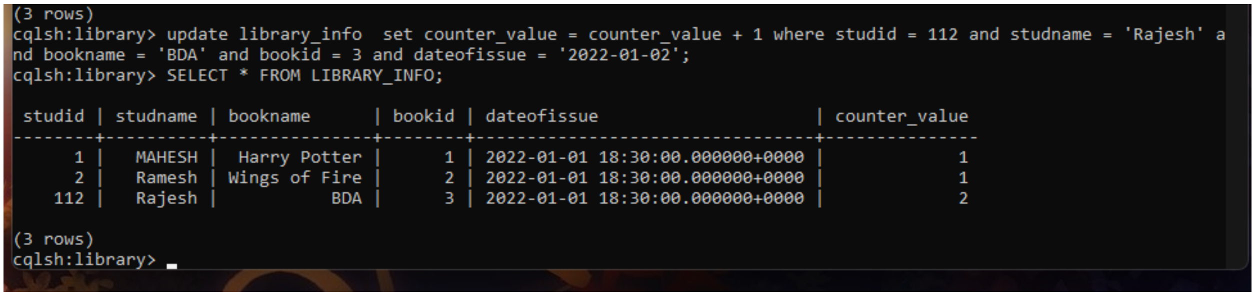


```
cqlsh:library> update library_info  set counter_value = counter_value + 1 where studid = 2 and studname = 'Ramesh' and
 bookname = 'Wings of Fire' and bookid = 2 and dateofissue = '2022-01-02';
cqlsh:library> SELECT * FROM LIBRARY_INFO;

 studid | studname | bookname      | bookid | dateofissue                     | counter_value
--------+----------+---------------+--------+---------------------------------+---------------
      1 |   MAHESH |  Harry Potter |      1 | 2022-01-01 18:30:00.000000+0000 |             1
      2 |   Ramesh | Wings of Fire |      2 | 2022-01-01 18:30:00.000000+0000 |             1

(2 rows)
cqlsh:library>
```

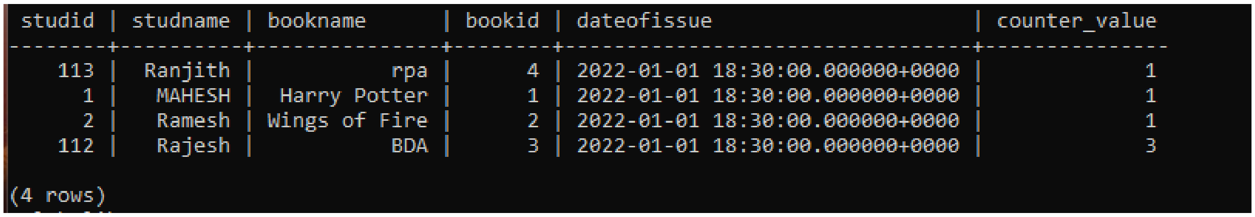
4. Display the details of the table created and increase the value of the counter



```
cqlsh:library> update library_info  set counter_value = counter_value + 1 where studid = 112 and studname = 'Rajesh' a
nd bookname = 'BDA' and bookid = 3 and dateofissue = '2022-01-02';
cqlsh:library> SELECT * FROM LIBRARY_INFO;

 studid | studname | bookname      | bookid | dateofissue                     | counter_value
--------+----------+---------------+--------+---------------------------------+---------------
      1 |   MAHESH |  Harry Potter |      1 | 2022-01-01 18:30:00.000000+0000 |             1
      2 |   Ramesh | Wings of Fire |      2 | 2022-01-01 18:30:00.000000+0000 |             1
    112 |   Rajesh |           BDA |      3 | 2022-01-01 18:30:00.000000+0000 |             1

(3 rows)
cqlsh:library>
```

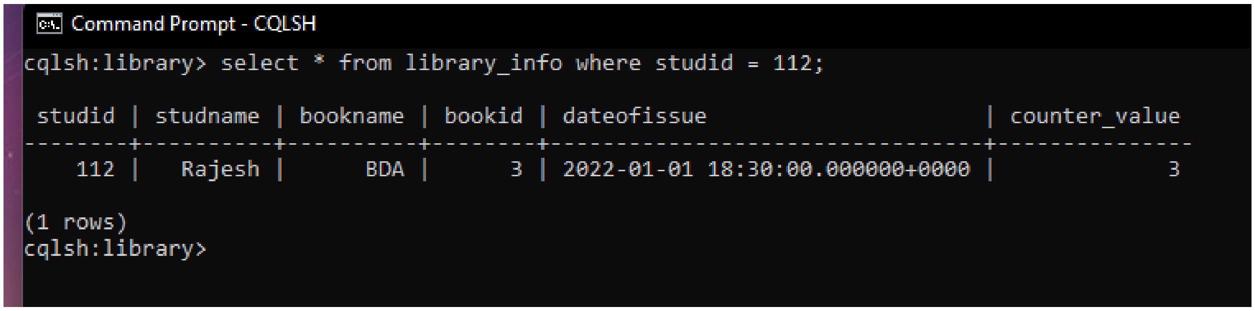


```
(3 rows)
cqlsh:library> update library_info  set counter_value = counter_value + 1 where studid = 112 and studname = 'Rajesh' a
nd bookname = 'BDA' and bookid = 3 and dateofissue = '2022-01-02';
cqlsh:library> SELECT * FROM LIBRARY_INFO;

 studid | studname | bookname      | bookid | dateofissue                     | counter_value
--------+----------+---------------+--------+---------------------------------+---------------
      1 |   MAHESH |  Harry Potter |      1 | 2022-01-01 18:30:00.000000+0000 |             1
      2 |   Ramesh | Wings of Fire |      2 | 2022-01-01 18:30:00.000000+0000 |             1
    112 |   Rajesh |           BDA |      3 | 2022-01-01 18:30:00.000000+0000 |             2

(3 rows)
cqlsh:library>
```



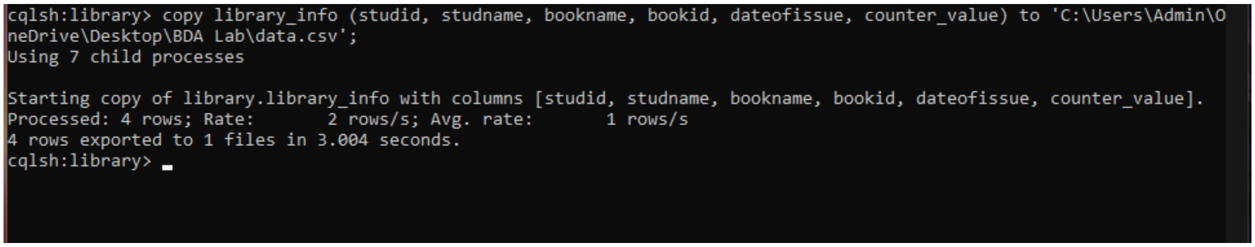
```
 studid | studname | bookname      | bookid | dateofissue                     | counter_value
--------+----------+---------------+--------+---------------------------------+---------------
    113 |  Ranjith |           rpa |      4 | 2022-01-01 18:30:00.000000+0000 |             1
      1 |   MAHESH |  Harry Potter |      1 | 2022-01-01 18:30:00.000000+0000 |             1
      2 |   Ramesh | Wings of Fire |      2 | 2022-01-01 18:30:00.000000+0000 |             1
    112 |   Rajesh |           BDA |      3 | 2022-01-01 18:30:00.000000+0000 |             3

(4 rows)
```

5. Write a query to show that a student with id 112 has taken a book “BDA” 3 times.

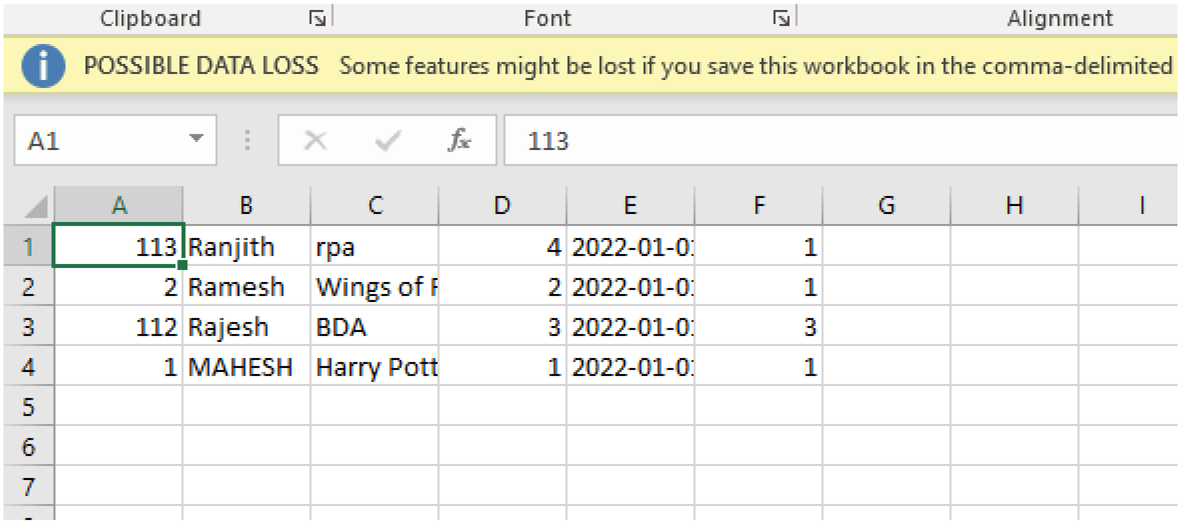


```
cqlsh:library> select * from library_info where studid = 112;

 studid | studname | bookname | bookid | dateofissue                     | counter_value
--------+----------+----------+--------+---------------------------------+---------------
    112 |   Rajesh |      BDA |      3 | 2022-01-01 18:30:00.000000+0000 |             3

(1 rows)
cqlsh:library>
```

6. Export the created column to a csv file



```
cqlsh:library> copy library_info (studid, studname, bookname, bookid, dateofissue, counter_value) to 'C:\Users\Admin\O
neDrive\Desktop\BDA Lab\data.csv';
Using 7 child processes

Starting copy of library.library_info with columns [studid, studname, bookname, bookid, dateofissue, counter_value].
Processed: 4 rows; Rate:       2 rows/s; Avg. rate:       1 rows/s
4 rows exported to 1 files in 3.004 seconds.
cqlsh:library>
```

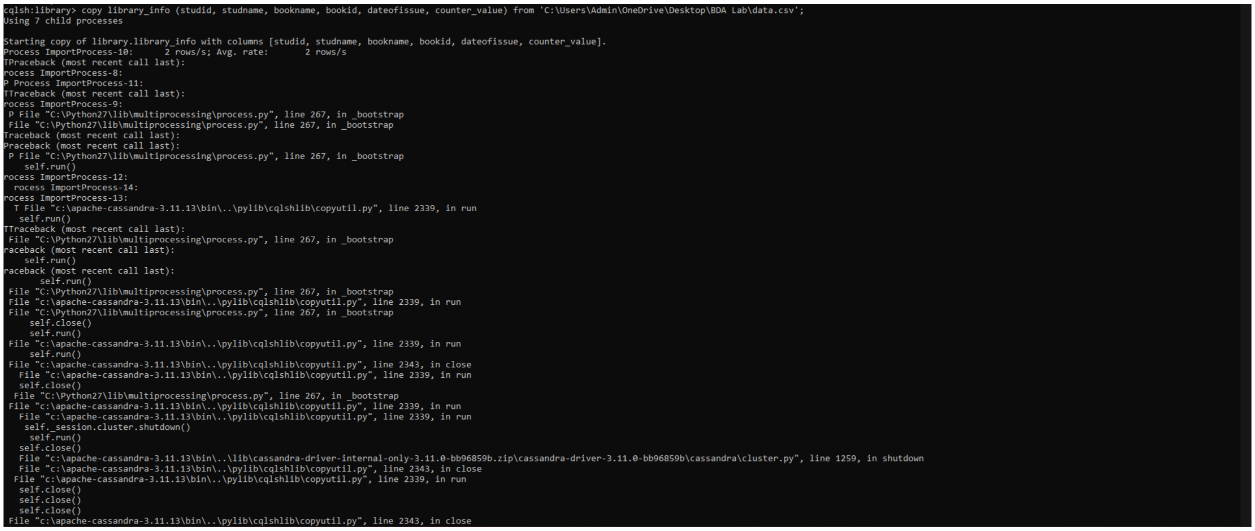


POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited

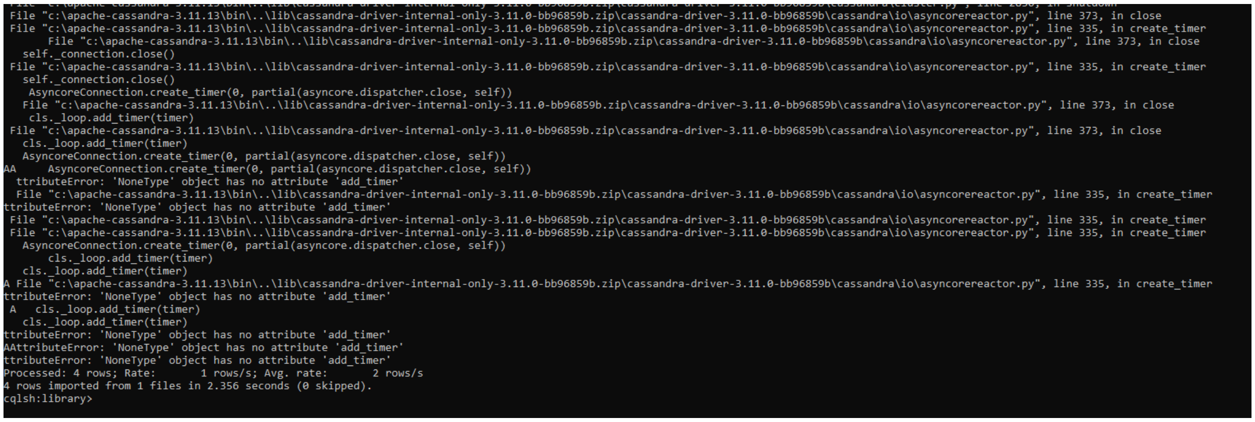
A1 | 113

	A	B	C	D	E	F	G	H	I
1	113	Ranjith	rpa	4	2022-01-0	1			
2	2	Ramesh	Wings of F	2	2022-01-0	1			
3	112	Rajesh	BDA	3	2022-01-0	3			
4	1	MAHESH	Harry Pott	1	2022-01-0	1			
5									
6									
7									

7. Import a given csv dataset from local file system into Cassandra column family



```
cqlsh:library> copy library_info (studid, studname, bookname, bookid, dateofissue, counter_value) from 'C:\Users\Admin\OneDrive\Desktop\BDA Lab\data.csv';
Using 7 child processes

Starting copy of library.library_info with columns [studid, studname, bookname, bookid, dateofissue, counter_value].
Process ImportProcess-10:      2 rows/s; Avg. rate:       2 rows/s
TPraceback (most recent call last):
rocess ImportProcess-8:
P Process ImportProcess-11:
TTraceback (most recent call last):
rocess ImportProcess-9:
 P File "C:\Python27\lib\multiprocessing\process.py", line 267, in _bootstrap
 File "C:\Python27\lib\multiprocessing\process.py", line 267, in _bootstrap
Traceback (most recent call last):
Praceback (most recent call last):
 P File "C:\Python27\lib\multiprocessing\process.py", line 267, in _bootstrap
    self.run()
rocess ImportProcess-12:
  rocess ImportProcess-14:
rocess ImportProcess-13:
  T File "c:\apache-cassandra-3.11.13\bin\..\pylib\cqlshlib\copyutil.py", line 2339, in run
   self.run()
TTraceback (most recent call last):
 File "C:\Python27\lib\multiprocessing\process.py", line 267, in _bootstrap
raceback (most recent call last):
    self.run()
raceback (most recent call last):
       self.run()
 File "C:\Python27\lib\multiprocessing\process.py", line 267, in _bootstrap
 File "c:\apache-cassandra-3.11.13\bin\..\pylib\cqlshlib\copyutil.py", line 2339, in run
 File "C:\Python27\lib\multiprocessing\process.py", line 267, in _bootstrap
     self.close()
     self.run()
 File "c:\apache-cassandra-3.11.13\bin\..\pylib\cqlshlib\copyutil.py", line 2339, in run
     self.run()
 File "c:\apache-cassandra-3.11.13\bin\..\pylib\cqlshlib\copyutil.py", line 2343, in close
   File "c:\apache-cassandra-3.11.13\bin\..\pylib\cqlshlib\copyutil.py", line 2339, in run
   self.close()
  File "C:\Python27\lib\multiprocessing\process.py", line 267, in _bootstrap
 File "c:\apache-cassandra-3.11.13\bin\..\pylib\cqlshlib\copyutil.py", line 2339, in run
   File "c:\apache-cassandra-3.11.13\bin\..\pylib\cqlshlib\copyutil.py", line 2339, in run
    self._session.cluster.shutdown()
     self.run()
   self.close()
   File "c:\apache-cassandra-3.11.13\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\cluster.py", line 1259, in shutdown
   File "c:\apache-cassandra-3.11.13\bin\..\pylib\cqlshlib\copyutil.py", line 2343, in close
  File "c:\apache-cassandra-3.11.13\bin\..\pylib\cqlshlib\copyutil.py", line 2339, in run
   self.close()
   self.close()
   self.close()
 File "c:\apache-cassandra-3.11.13\bin\..\pylib\cqlshlib\copyutil.py", line 2343, in close
```



```
 File "c:\apache-cassandra-3.11.13\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\cluster.py", line 2836, in shutdown
 File "c:\apache-cassandra-3.11.13\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\io\asyncorereactor.py", line 373, in close
 File "c:\apache-cassandra-3.11.13\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\io\asyncorereactor.py", line 335, in create_timer
       File "c:\apache-cassandra-3.11.13\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\io\asyncorereactor.py", line 373, in close
   self._connection.close()
 File "c:\apache-cassandra-3.11.13\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\io\asyncorereactor.py", line 335, in create_timer
   self._connection.close()
    AsyncoreConnection.create_timer(0, partial(asyncore.dispatcher.close, self))
   File "c:\apache-cassandra-3.11.13\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\io\asyncorereactor.py", line 373, in close
    cls._loop.add_timer(timer)
 File "c:\apache-cassandra-3.11.13\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\io\asyncorereactor.py", line 373, in close
   cls._loop.add_timer(timer)
   AsyncoreConnection.create_timer(0, partial(asyncore.dispatcher.close, self))
AA     AsyncoreConnection.create_timer(0, partial(asyncore.dispatcher.close, self))
  ttributeError: 'NoneType' object has no attribute 'add_timer'
  File "c:\apache-cassandra-3.11.13\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\io\asyncorereactor.py", line 335, in create_timer
ttributeError: 'NoneType' object has no attribute 'add_timer'
 File "c:\apache-cassandra-3.11.13\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\io\asyncorereactor.py", line 335, in create_timer
 File "c:\apache-cassandra-3.11.13\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\io\asyncorereactor.py", line 335, in create_timer
   AsyncoreConnection.create_timer(0, partial(asyncore.dispatcher.close, self))
       cls._loop.add_timer(timer)
   cls._loop.add_timer(timer)
A File "c:\apache-cassandra-3.11.13\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\io\asyncorereactor.py", line 335, in create_timer
ttributeError: 'NoneType' object has no attribute 'add_timer'
 A   cls._loop.add_timer(timer)
   cls._loop.add_timer(timer)
ttributeError: 'NoneType' object has no attribute 'add_timer'
AAttributeError: 'NoneType' object has no attribute 'add_timer'
ttributeError: 'NoneType' object has no attribute 'add_timer'
Processed: 4 rows; Rate:       1 rows/s; Avg. rate:       2 rows/s
4 rows imported from 1 files in 2.356 seconds (0 skipped).
cqlsh:library>
```

MongoDB Lab Program 1 (CRUD Demonstration): -

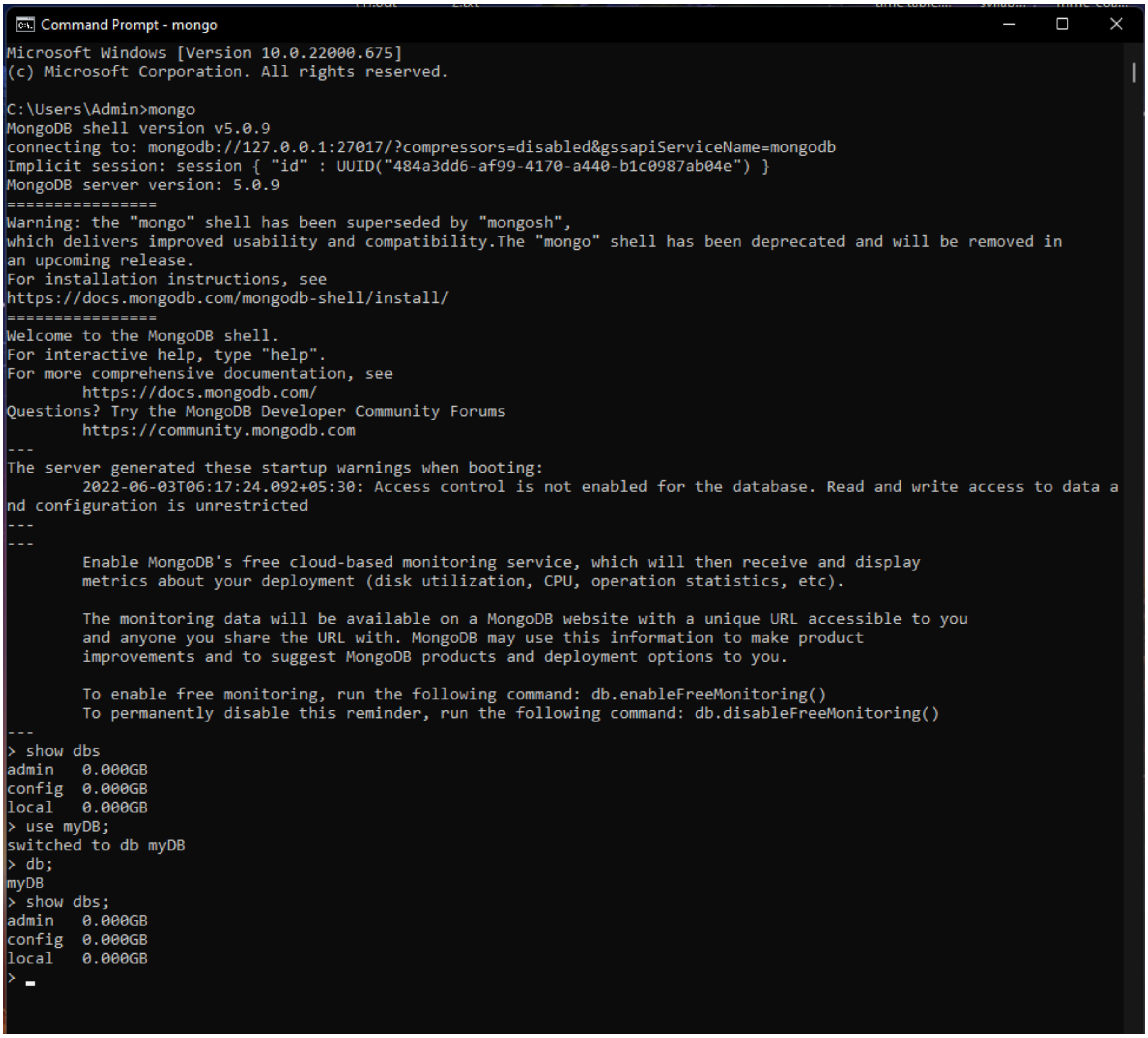
Execute the queries and upload a document with output.

I. CREATE DATABASE IN MONGODB.

use myDB;

db;   (Confirm the existence of your database)

show dbs; (To list all databases)



```
Microsoft Windows [Version 10.0.22000.675]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Admin>mongo
MongoDB shell version v5.0.9
connecting to: mongodb://127.0.0.1:27017/?compressors=disabled&gssapiServiceName=mongodb
Implicit session: session { "id" : UUID("484a3dd6-af99-4170-a440-b1c0987ab04e") }
MongoDB server version: 5.0.9
================
Warning: the "mongo" shell has been superseded by "mongosh",
which delivers improved usability and compatibility.The "mongo" shell has been deprecated and will be removed in
an upcoming release.
For installation instructions, see
https://docs.mongodb.com/mongodb-shell/install/
================
Welcome to the MongoDB shell.
For interactive help, type "help".
For more comprehensive documentation, see
        https://docs.mongodb.com/
Questions? Try the MongoDB Developer Community Forums
        https://community.mongodb.com
---
The server generated these startup warnings when booting:
        2022-06-03T06:17:24.092+05:30: Access control is not enabled for the database. Read and write access to data a
nd configuration is unrestricted
---
---
        Enable MongoDB's free cloud-based monitoring service, which will then receive and display
        metrics about your deployment (disk utilization, CPU, operation statistics, etc).

        The monitoring data will be available on a MongoDB website with a unique URL accessible to you
        and anyone you share the URL with. MongoDB may use this information to make product
        improvements and to suggest MongoDB products and deployment options to you.

        To enable free monitoring, run the following command: db.enableFreeMonitoring()
        To permanently disable this reminder, run the following command: db.disableFreeMonitoring()
---
> show dbs
admin   0.000GB
config  0.000GB
local   0.000GB
> use myDB;
switched to db myDB
> db;
myDB
> show dbs;
admin   0.000GB
config  0.000GB
local   0.000GB
> 
```

II.CRUD (CREATE, READ, UPDATE, DELETE) OPERATIONS

1. To create a collection by the name “Student”. Let us take a look at the collection list

prior to the creation of the new collection “Student”.

db.createCollection(“Student”); =&gt; sql equivalent CREATE TABLE STUDENT(…);

2. To drop a collection by the name “Student”.

db.Student.drop();

3. Create a collection by the name “Students” and store the following data in it.

db.Student.insert({_id:1,StudName:&quot;MichelleJacintha&quot;,Grade:&quot;VII&quot;,Hobbies:&quot;InternetS

urfing&quot;});

4. Insert the document for “AryanDavid” in to the Students collection only if it does not

already exist in the collection. However, if it is already present in the collection, then

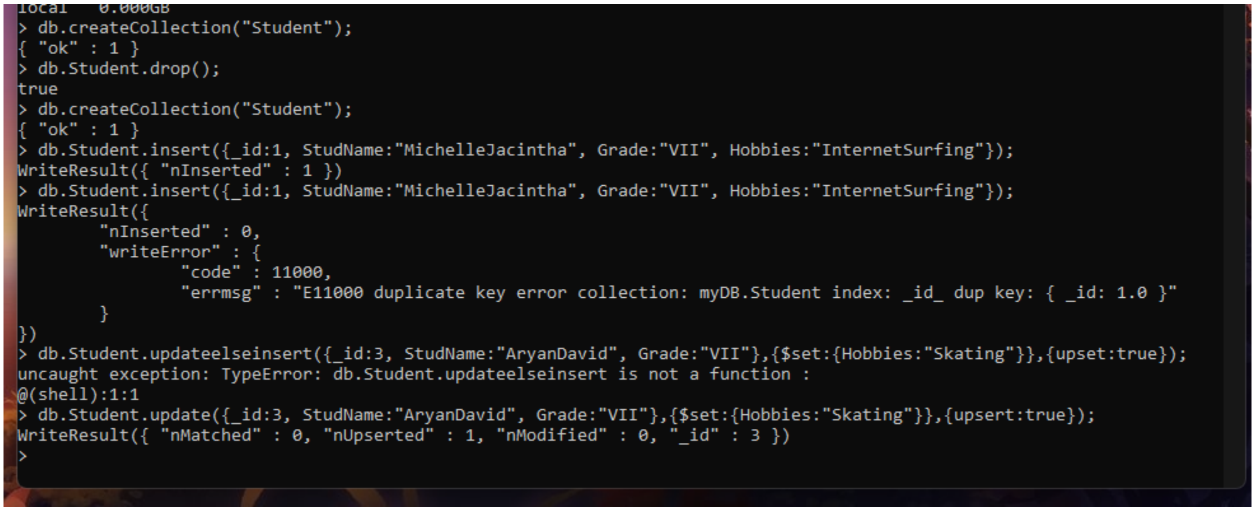
update the document with new values. (Update his Hobbies from “Skating” to “Chess”.

) Use “Update else insert” (if there is an existing document, it will attempt to update it,

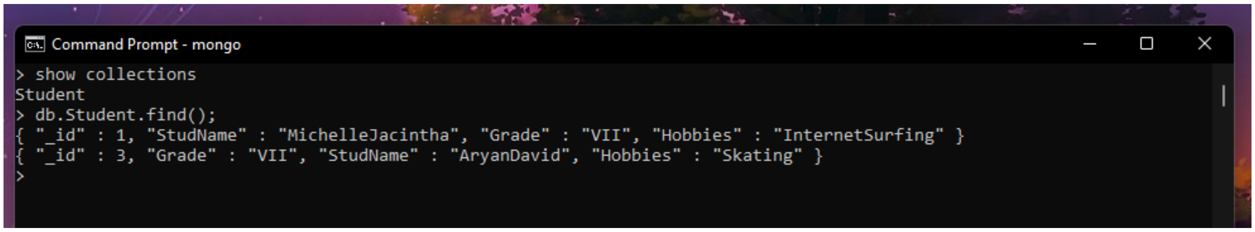
if there is no existing document then it will insert it).

db.Student.update({_id:3,StudName:&quot;AryanDavid&quot;,Grade:&quot;VII&quot;},{$set:{Hobbies:&quot;Skatin

g&quot;}},{upsert:true});



```
local    0.000GB
> db.createCollection("Student");
{ "ok" : 1 }
> db.Student.drop();
true
> db.createCollection("Student");
{ "ok" : 1 }
> db.Student.insert({_id:1, StudName:"MichelleJacintha", Grade:"VII", Hobbies:"InternetSurfing"});
WriteResult({ "nInserted" : 1 })
> db.Student.insert({_id:1, StudName:"MichelleJacintha", Grade:"VII", Hobbies:"InternetSurfing"});
WriteResult({
        "nInserted" : 0,
        "writeError" : {
                "code" : 11000,
                "errmsg" : "E11000 duplicate key error collection: myDB.Student index: _id_ dup key: { _id: 1.0 }"
        }
})
> db.Student.updateelseinsert({_id:3, StudName:"AryanDavid", Grade:"VII"},{$set:{Hobbies:"Skating"}},{upset:true});
uncaught exception: TypeError: db.Student.updateelseinsert is not a function :
@(shell):1:1
> db.Student.update({_id:3, StudName:"AryanDavid", Grade:"VII"},{$set:{Hobbies:"Skating"}},{upsert:true});
WriteResult({ "nMatched" : 0, "nUpserted" : 1, "nModified" : 0, "_id" : 3 })
>
```



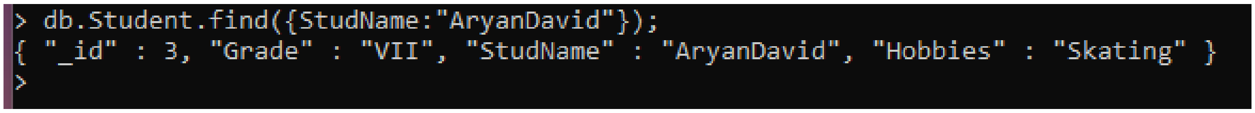
```
> show collections
Student
> db.Student.find();
{ "_id" : 1, "StudName" : "MichelleJacintha", "Grade" : "VII", "Hobbies" : "InternetSurfing" }
{ "_id" : 3, "Grade" : "VII", "StudName" : "AryanDavid", "Hobbies" : "Skating" }
>
```

5. FIND METHOD

A. To search for documents from the “Students” collection based on certain search criteria.

db.Student.find({StudName:&quot;Aryan David&quot;});

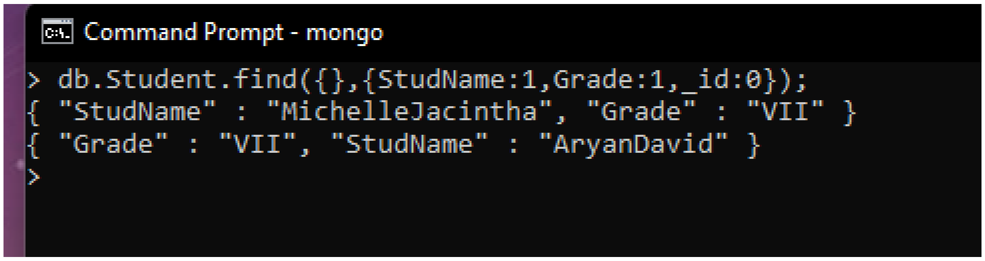
({cond..},{columns.. column:1, columnname:0})



```
> db.Student.find({StudName:"AryanDavid"});
{ "_id" : 3, "Grade" : "VII", "StudName" : "AryanDavid", "Hobbies" : "Skating" }
>
```

B. To display only the StudName and Grade from all the documents of the Students collection. The identifier_id should be suppressed and NOT displayed.

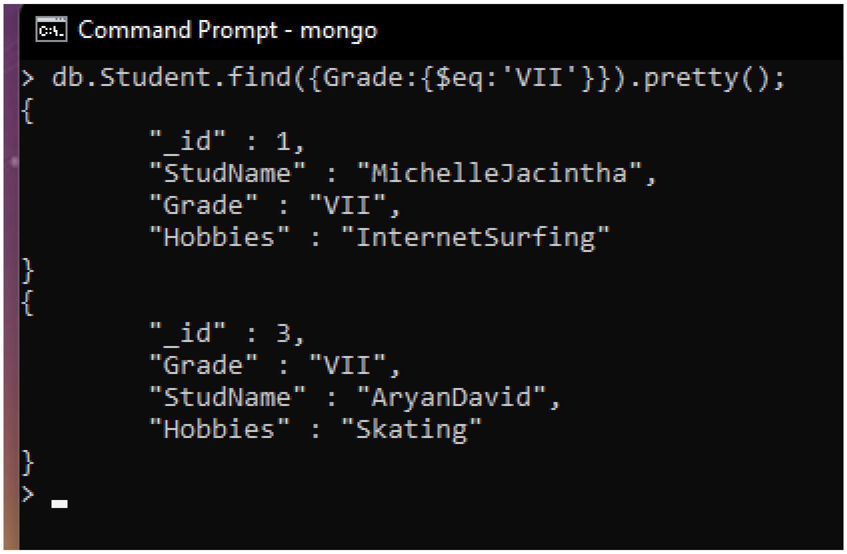
db.Student.find({},{StudName:1,Grade:1,_id:0});



```
> db.Student.find({},{StudName:1,Grade:1,_id:0});
{ "StudName" : "MichelleJacintha", "Grade" : "VII" }
{ "Grade" : "VII", "StudName" : "AryanDavid" }
>
```

C. To find those documents where the Grade is set to ‘VII’

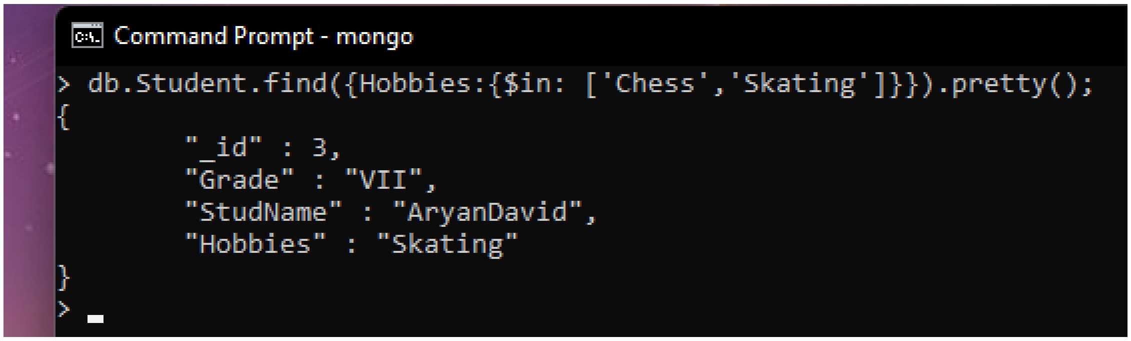
db.Student.find({Grade:{$eq:&#39;VII&#39;}}).pretty();



```
> db.Student.find({Grade:{$eq:'VII'}}).pretty();
{
        "_id" : 1,
        "StudName" : "MichelleJacintha",
        "Grade" : "VII",
        "Hobbies" : "InternetSurfing"
}
{
        "_id" : 3,
        "Grade" : "VII",
        "StudName" : "AryanDavid",
        "Hobbies" : "Skating"
}
> 
```

D. To find those documents from the Students collection where the Hobbies is set to either ‘Chess’ or is set to ‘Skating’.

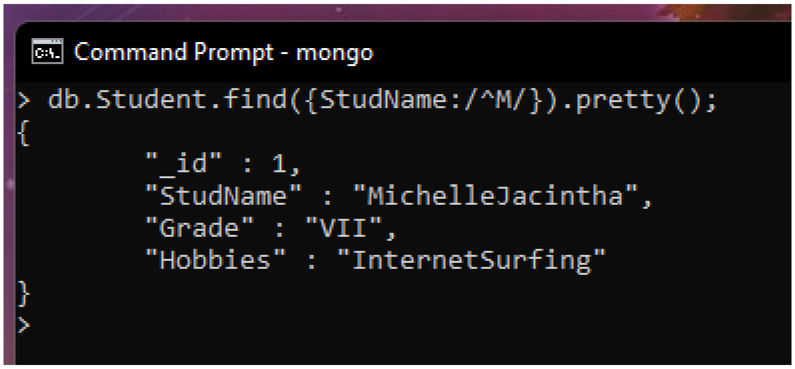
db.Student.find({Hobbies :{ $in: [&#39;Chess&#39;,&#39;Skating&#39;]}}).pretty ();



```
> db.Student.find({Hobbies:{$in: ['Chess','Skating']}}).pretty();
{
        "_id" : 3,
        "Grade" : "VII",
        "StudName" : "AryanDavid",
        "Hobbies" : "Skating"
}
> 
```

E. To find documents from the Students collection where the StudName begins with “M”.

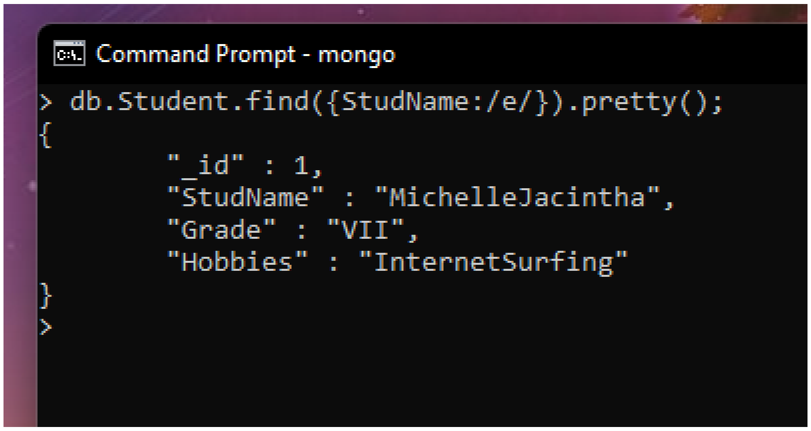
db.Student.find({StudName:/^M/}).pretty();



```
> db.Student.find({StudName:/^M/}).pretty();
{
        "_id" : 1,
        "StudName" : "MichelleJacintha",
        "Grade" : "VII",
        "Hobbies" : "InternetSurfing"
}
>
```

F. To find documents from the Students collection where the StudNamehas an “e” in any position.

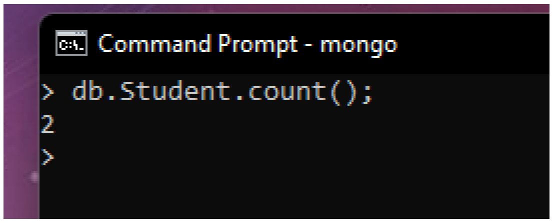
db.Student.find({StudName:/e/}).pretty();



```
> db.Student.find({StudName:/e/}).pretty();
{
        "_id" : 1,
        "StudName" : "MichelleJacintha",
        "Grade" : "VII",
        "Hobbies" : "InternetSurfing"
}
>
```

G. To find the number of documents in the Students collection.

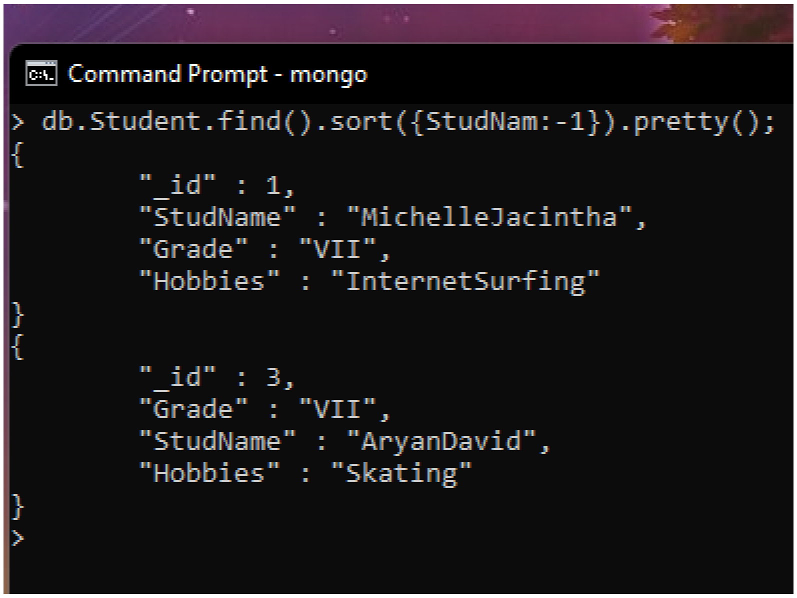
db.Student.count();



```
> db.Student.count();
2
>
```

H. To sort the documents from the Students collection in the descending order of StudName.

db.Student.find().sort({StudName:-1}).pretty();

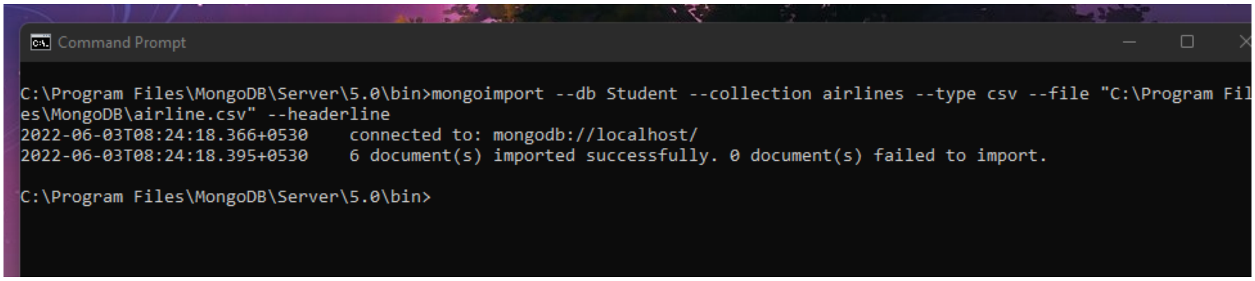


```
> db.Student.find().sort({StudNam:-1}).pretty();
{
        "_id" : 1,
        "StudName" : "MichelleJacintha",
        "Grade" : "VII",
        "Hobbies" : "InternetSurfing"
}
{
        "_id" : 3,
        "Grade" : "VII",
        "StudName" : "AryanDavid",
        "Hobbies" : "Skating"
}
>
```

III. Import data from a CSV file

Given a CSV file “sample.txt” in the D:drive, import the file into the MongoDB collection, “SampleJSON”. The collection is in the database “test”.

mongoimport --db Student --collection airlines --type csv –headerline --file /home/hduser/Desktop/airline.csv

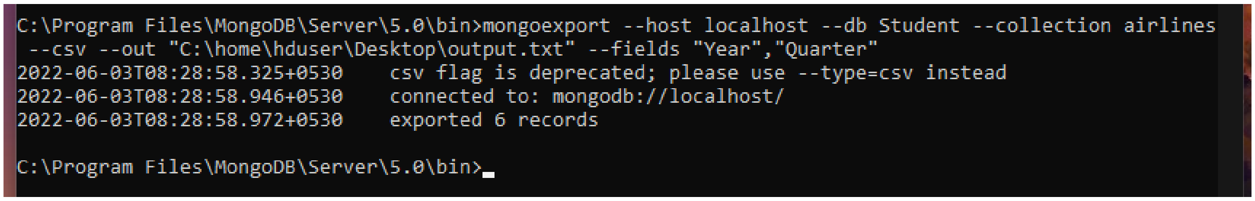


```
C:\Program Files\MongoDB\Server\5.0\bin>mongoimport --db Student --collection airlines --type csv --file "C:\Program Fil
es\MongoDB\airline.csv" --headerline
2022-06-03T08:24:18.366+0530    connected to: mongodb://localhost/
2022-06-03T08:24:18.395+0530    6 document(s) imported successfully. 0 document(s) failed to import.

C:\Program Files\MongoDB\Server\5.0\bin>
```

IV. Export data to a CSV file

This command used at the command prompt exports MongoDB JSON documents from “Customers” collection in the “test” database into a CSV file “Output.txt” in the D:drive.

mongoexport --host localhost --db Student --collection airlines --csv --out /home/hduser/Desktop/output.txt –fields “Year”,”Quarter”

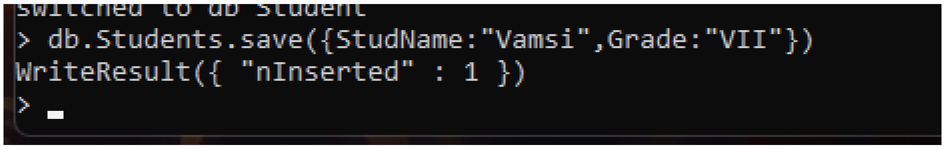


```
C:\Program Files\MongoDB\Server\5.0\bin>mongoexport --host localhost --db Student --collection airlines
 --csv --out "C:\home\hduser\Desktop\output.txt" --fields "Year","Quarter"
2022-06-03T08:28:58.325+0530    csv flag is deprecated; please use --type=csv instead
2022-06-03T08:28:58.946+0530    connected to: mongodb://localhost/
2022-06-03T08:28:58.972+0530    exported 6 records

C:\Program Files\MongoDB\Server\5.0\bin>
```

V. Save Method :

Save() method will insert a new document, if the document with the _id does not exist. If it exists it will replace the exisiting document.

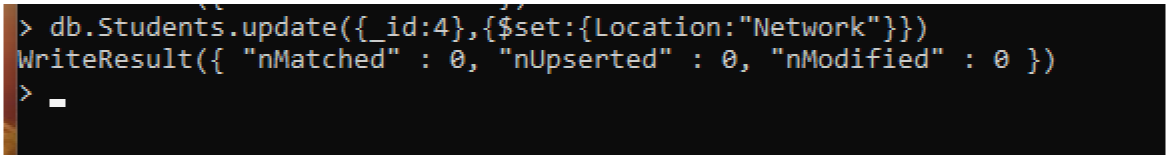
db.Students.save({StudName:”Vamsi”, Grade:”VI”})



```
switched to db Student
> db.Students.save({StudName:"Vamsi",Grade:"VII"})
WriteResult({ "nInserted" : 1 })
>
```

VI. Add a new field to existing Document:

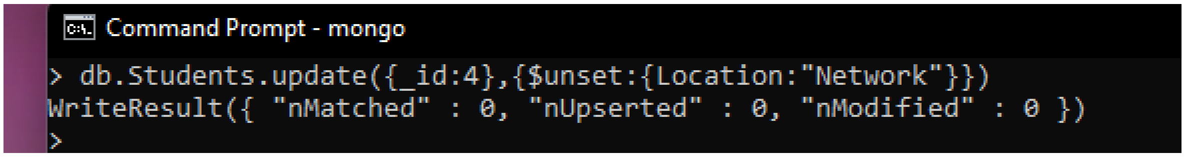
db.Students.update({_id:4},{$set:{Location:”Network”}})



```
> db.Students.update({_id:4},{$set:{Location:"Network"}})
WriteResult({ "nMatched" : 0, "nUpserted" : 0, "nModified" : 0 })
>
```

VII. Remove the field in an existing Document

db.Students.update({_id:4},{$unset:{Location:”Network”}})



```
> db.Students.update({_id:4},{$unset:{Location:"Network"}})
WriteResult({ "nMatched" : 0, "nUpserted" : 0, "nModified" : 0 })
>
```

VIII. Finding Document based on search criteria suppressing few fields

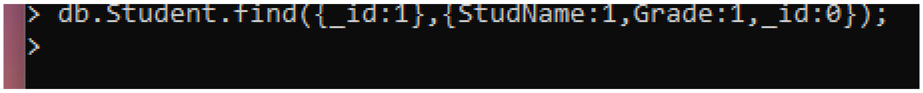
db.Student.find({_id:1},{StudName:1,Grade:1,_id:0});

To find those documents where the Grade is not set to ‘VII’

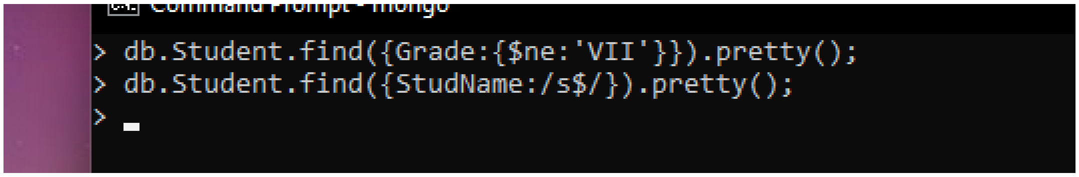
db.Student.find({Grade:{$ne:&#39;VII&#39;}}).pretty();

To find documents from the Students collection where the StudName ends with s.

db.Student.find({StudName:/s$/}).pretty();

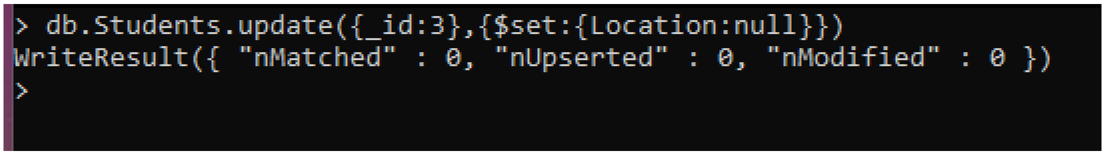


```
> db.Student.find({_id:1},{StudName:1,Grade:1,_id:0});
>
```



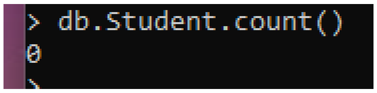
```
> db.Student.find({Grade:{$ne:'VII'}}).pretty();
> db.Student.find({StudName:/s$/}).pretty();
> 
```

IX. to set a particular field value to NULL



```
> db.Students.update({_id:3},{$set:{Location:null}})
WriteResult({ "nMatched" : 0, "nUpserted" : 0, "nModified" : 0 })
>
```

X Count the number of documents in Student Collections



```
> db.Student.count()
0
```

XI. Count the number of documents in Student Collections with grade :VII

db.Students.count({Grade:”VII”})

retrieve first 3 documents

db.Students.find({Grade:”VII”}).limit(3).pretty();

Sort the document in Ascending order

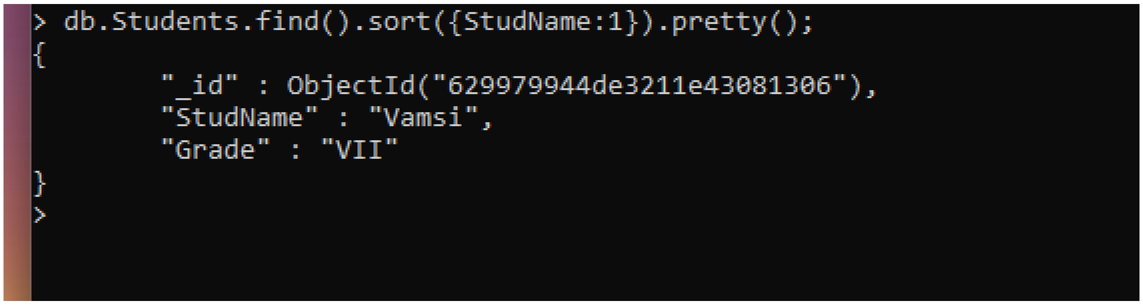
db.Students.find().sort({StudName:1}).pretty();

Note:

for desending order : db.Students.find().sort({StudName:-1}).pretty();

to Skip the 1 st two documents from the Students Collections

db.Students.find().skip(2).pretty()



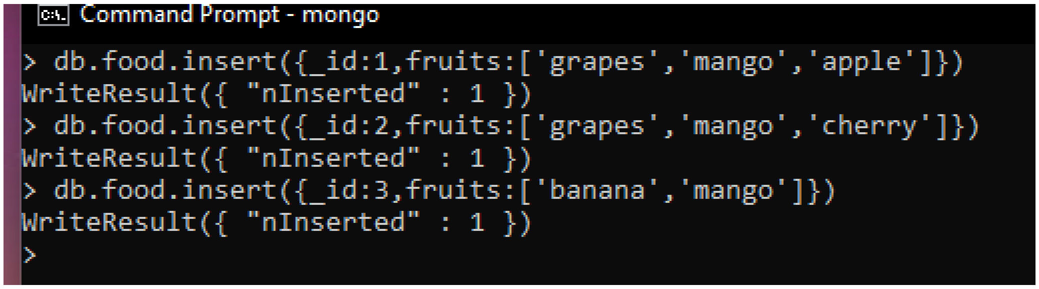
```
> db.Students.find().sort({StudName:1}).pretty();
{
        "_id" : ObjectId("629979944de3211e43081306"),
        "StudName" : "Vamsi",
        "Grade" : "VII"
}
>
```

XII. Create a collection by name “food” and add to each document add a “fruits” array

db.food.insert({ _id:1, fruits:[&#39;grapes&#39;,&#39;mango&#39;,&#39;apple&#39;] })

db.food.insert({ _id:2, fruits:[&#39;grapes&#39;,&#39;mango&#39;,&#39;cherry&#39;] })

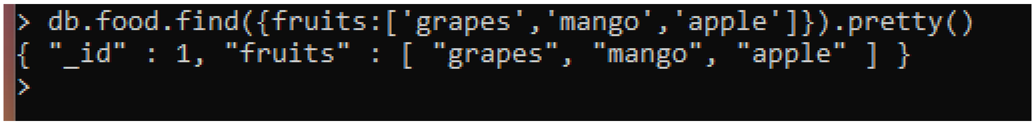
db.food.insert({ _id:3, fruits:[&#39;banana&#39;,&#39;mango&#39;] })



```
> db.food.insert({_id:1,fruits:['grapes','mango','apple']})
WriteResult({ "nInserted" : 1 })
> db.food.insert({_id:2,fruits:['grapes','mango','cherry']})
WriteResult({ "nInserted" : 1 })
> db.food.insert({_id:3,fruits:['banana','mango']})
WriteResult({ "nInserted" : 1 })
>
```

To find those documents from the “food” collection which has the “fruits array” constitute of “grapes”, “mango” and “apple”.

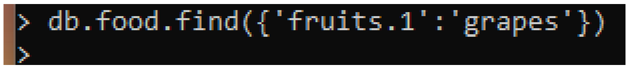
db.food.find ({fruits: [&#39;grapes&#39;,&#39;mango&#39;,&#39;apple&#39;] }). pretty().



```
> db.food.find({fruits:['grapes','mango','apple']}).pretty()
{ "_id" : 1, "fruits" : [ "grapes", "mango", "apple" ] }
>
```

To find in “fruits” array having “mango” in the first index position.

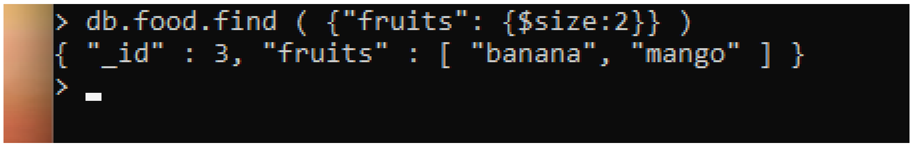
db.food.find ({&#39;fruits.1&#39;:&#39;grapes&#39;})



```
> db.food.find({'fruits.1':'grapes'})
>
```

To find those documents from the “food” collection where the size of the array is two.

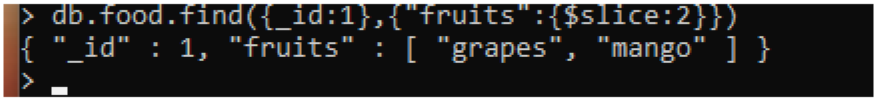
db.food.find ({“fruits”: {$size:2}})



```
> db.food.find ( {"fruits": {$size:2}} )
{ "_id" : 3, "fruits" : [ "banana", "mango" ] }
> 
```

To find the document with a particular id and display the first two elements from the array “fruits”

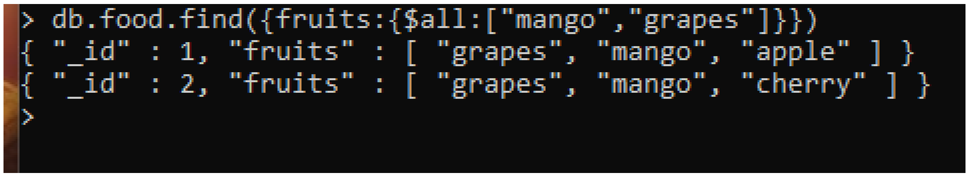
db.food.find({_id:1},{“fruits”:{$slice:2}})



```
> db.food.find({_id:1},{"fruits":{$slice:2}})
{ "_id" : 1, "fruits" : [ "grapes", "mango" ] }
> 
```

To find all the documets from the food collection which have elements mango and grapes in the array “fruits”

db.food.find({fruits:{$all:[“mango”,”grapes”]}})



```
> db.food.find({fruits:{$all:["mango","grapes"]}})
{ "_id" : 1, "fruits" : [ "grapes", "mango", "apple" ] }
{ "_id" : 2, "fruits" : [ "grapes", "mango", "cherry" ] }
>
```

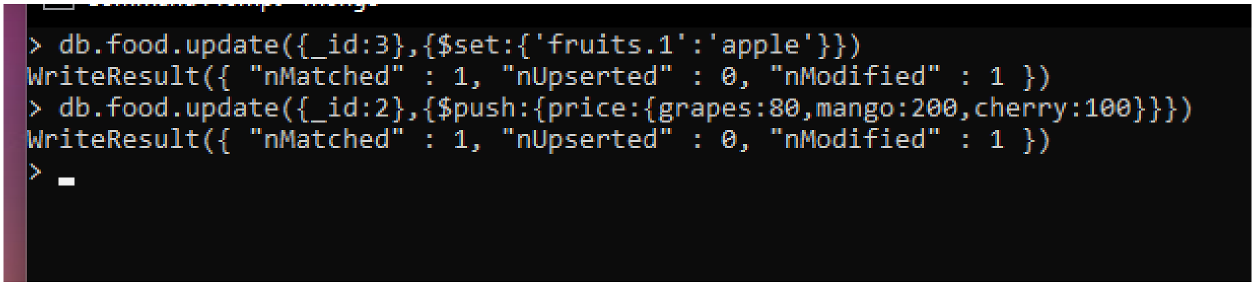
update on Array:

using particular id replace the element present in the 1 st index position of the fruits array with apple

db.food.update({_id:3},{$set:{&#39;fruits.1&#39;:&#39;apple&#39;}})

insert new key value pairs in the fruits array

db.food.update({_id:2},{$push:{price:{grapes:80,mango:200,cherry:100}}})



```
> db.food.update({_id:3},{$set:{'fruits.1':'apple'}})
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
> db.food.update({_id:2},{$push:{price:{grapes:80,mango:200,cherry:100}}})
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
> 
```

Note: perform query operations using - pop, addToSet, pullAll and pull

XII. Aggregate Function :

Create a collection Customers with fields custID, AcctBal, AcctType.

Now group on “custID” and compute the sum of “AccBal”.

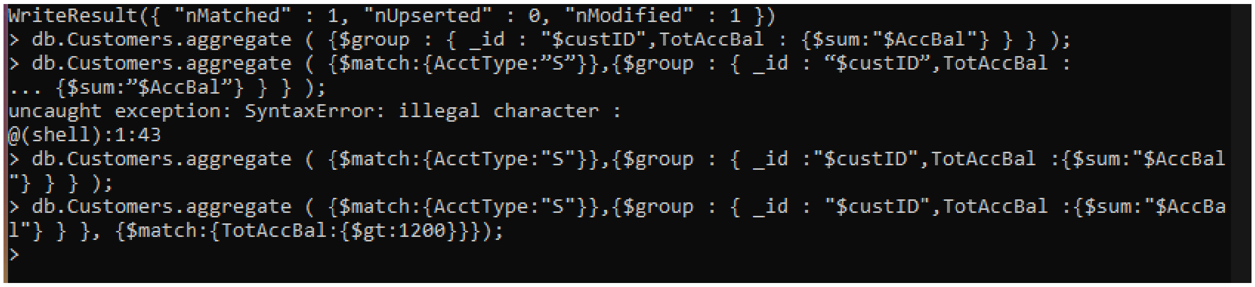
db.Customers.aggregate ({$group : { _id : “$custID”,TotAccBal : {$sum:”$AccBal”} } });

match on AcctType:”S” then group on “CustID” and compute the sum of “AccBal”.

db.Customers.aggregate ({$match:{AcctType:”S”}},{$group : { _id : “$custID”,TotAccBal : {$sum:”$AccBal”} } });

match on AcctType:”S” then group on “CustID” and compute the sum of “AccBal” and total balance greater than 1200.

db.Customers.aggregate ({$match:{AcctType:”S”}},{$group : { _id : “$custID”,TotAccBal : {$sum:”$AccBal”} } }, {$match:{TotAccBal:{$gt:1200}}});

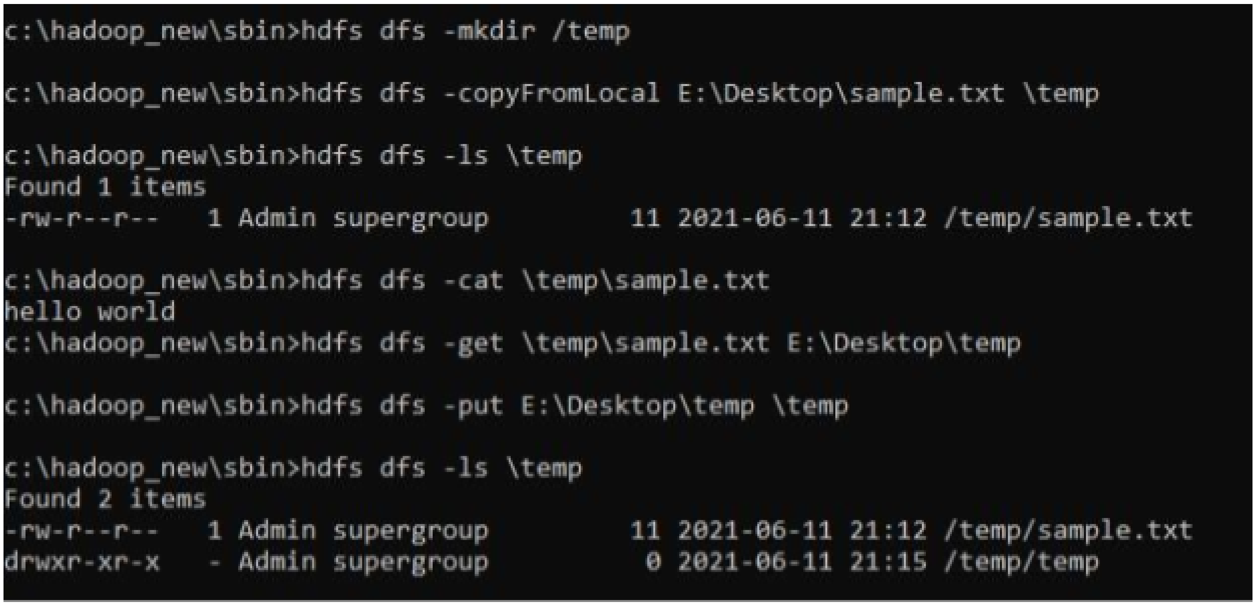


```
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
> db.Customers.aggregate ( {$group : { _id : "$custID",TotAccBal : {$sum:"$AccBal"} } } );
> db.Customers.aggregate ( {$match:{AcctType:”S”}},{$group : { _id : “$custID”,TotAccBal :
... {$sum:”$AccBal”} } } );
uncaught exception: SyntaxError: illegal character :
@(shell):1:43
> db.Customers.aggregate ( {$match:{AcctType:"S"}},{$group : { _id :"$custID",TotAccBal :{$sum:"$AccBal
"} } } );
> db.Customers.aggregate ( {$match:{AcctType:"S"}},{$group : { _id : "$custID",TotAccBal :{$sum:"$AccBa
l"} } }, {$match:{TotAccBal:{$gt:1200}}});
>
```

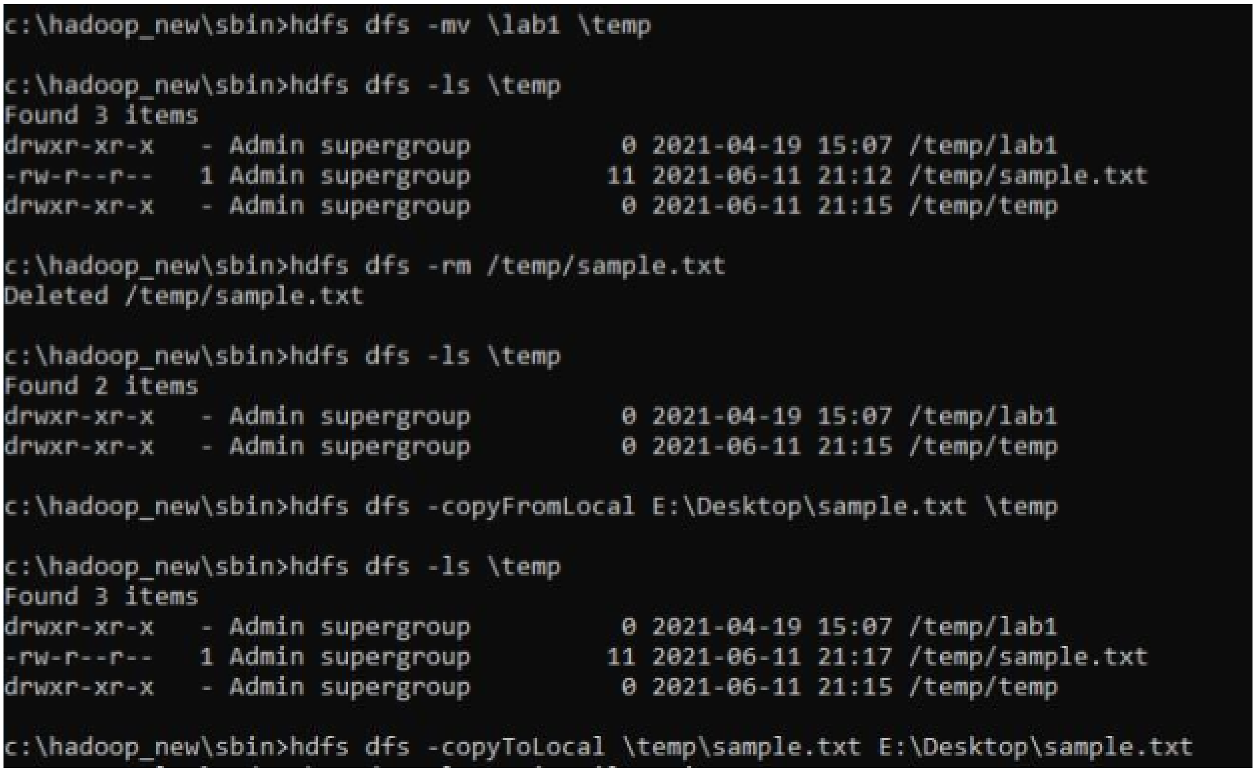
LAB 5

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

```
c:\hadoop_new\sbin>hdfs dfs -mkdir /temp


c:\hadoop_new\sbin>hdfs dfs -copyFromLocal E:\Desktop\sample.txt \temp


c:\hadoop_new\sbin>hdfs dfs -ls \temp
Found 1 items
-rw-r--r--   1 Admin supergroup        11 2021-06-11 21:12 /temp/sample.txt


c:\hadoop_new\sbin>hdfs dfs -cat \temp\sample.txt hello
world


c:\hadoop_new\sbin>hdfs dfs -get \temp\sample.txt E:\Desktop\temp


c:\hadoop_new\sbin>hdfs dfs -put E:\Desktop\temp \temp


c:\hadoop_new\sbin>hdfs dfs -ls \temp
Found 2 items
-rw-r--r--   1 Admin supergroup        11 2021-06-11 21:12 /temp/sample.txt drwxr-xr-x   -
Admin supergroup         0 2021-06-11 21:15 /temp/temp


c:\hadoop_new\sbin>hdfs dfs -mv \lab1 \temp


c:\hadoop_new\sbin>hdfs dfs -ls \temp Found 3 items drwxr-xr-x   - Admin
supergroup         0 2021-04-19 15:07 /temp/lab1 -rw-r--r--   1 Admin
supergroup        11 2021-06-11 21:12 /temp/sample.txt drwxr-xr-x   -
Admin supergroup         0 2021-06-11 21:15 /temp/temp

c:\hadoop_new\sbin>hdfs dfs -rm /temp/sample.txt
Deleted /temp/sample.txt

c:\hadoop_new\sbin>hdfs dfs -ls \temp Found 2 items drwxr-xr-x   - Admin
supergroup         0 2021-04-19 15:07 /temp/lab1 drwxr-xr-x   - Admin
supergroup         0 2021-06-11 21:15 /temp/temp

c:\hadoop_new\sbin>hdfs dfs -copyFromLocal E:\Desktop\sample.txt \temp

c:\hadoop_new\sbin>hdfs dfs -ls \temp Found 3 items drwxr-xr-x   - Admin
supergroup         0 2021-04-19 15:07 /temp/lab1 -rw-r--r--   1 Admin supergroup
11 2021-06-11 21:17 /temp/sample.txt drwxr-xr-x   - Admin supergroup         0
2021-06-11 21:15 /temp/temp

c:\hadoop_new\sbin>hdfs dfs -copyToLocal \temp\sample.txt E:\Desktop\sample.txt
```



```
c:\hadoop_new\sbin>hdfs dfs -mkdir /temp

c:\hadoop_new\sbin>hdfs dfs -copyFromLocal E:\Desktop\sample.txt \temp

c:\hadoop_new\sbin>hdfs dfs -ls \temp
Found 1 items
-rw-r--r--   1 Admin supergroup         11 2021-06-11 21:12 /temp/sample.txt

c:\hadoop_new\sbin>hdfs dfs -cat \temp\sample.txt
hello world
c:\hadoop_new\sbin>hdfs dfs -get \temp\sample.txt E:\Desktop\temp

c:\hadoop_new\sbin>hdfs dfs -put E:\Desktop\temp \temp

c:\hadoop_new\sbin>hdfs dfs -ls \temp
Found 2 items
-rw-r--r--   1 Admin supergroup         11 2021-06-11 21:12 /temp/sample.txt
drwxr-xr-x   - Admin supergroup          0 2021-06-11 21:15 /temp/temp
```

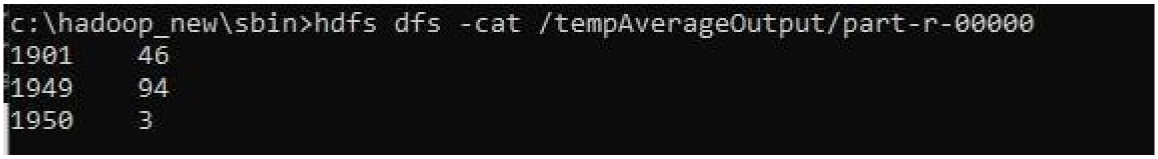


```
c:\hadoop_new\sbin>hdfs dfs -mv \lab1 \temp

c:\hadoop_new\sbin>hdfs dfs -ls \temp
Found 3 items
drwxr-xr-x   - Admin supergroup          0 2021-04-19 15:07 /temp/lab1
-rw-r--r--   1 Admin supergroup         11 2021-06-11 21:12 /temp/sample.txt
drwxr-xr-x   - Admin supergroup          0 2021-06-11 21:15 /temp/temp

c:\hadoop_new\sbin>hdfs dfs -rm /temp/sample.txt
Deleted /temp/sample.txt

c:\hadoop_new\sbin>hdfs dfs -ls \temp
Found 2 items
drwxr-xr-x   - Admin supergroup          0 2021-04-19 15:07 /temp/lab1
drwxr-xr-x   - Admin supergroup          0 2021-06-11 21:15 /temp/temp

c:\hadoop_new\sbin>hdfs dfs -copyFromLocal E:\Desktop\sample.txt \temp

c:\hadoop_new\sbin>hdfs dfs -ls \temp
Found 3 items
drwxr-xr-x   - Admin supergroup          0 2021-04-19 15:07 /temp/lab1
-rw-r--r--   1 Admin supergroup         11 2021-06-11 21:17 /temp/sample.txt
drwxr-xr-x   - Admin supergroup          0 2021-06-11 21:15 /temp/temp

c:\hadoop_new\sbin>hdfs dfs -copyToLocal \temp\sample.txt E:\Desktop\sample.txt
```

LAB 6

For the given file, Create a Map Reduce program to
a) Find the average temperature for each year from the NCDC data set.

```
// AverageDriver.java package temperature;

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

public class AverageDriver
{           public static void main (String[] args) throws Exception
            {
                        if (args.length != 2)
                        {
                                    System.err.println("Please Enter the input and output parameters");
                                    System.exit(-1);
                        }
                        Job job = new Job();                    job.setJarByClass(AverageDriver.class);
job.setJobName("Max temperature");
                        FileInputFormat.addInputPath(job,new Path(args[0]));
                        FileOutputFormat.setOutputPath(job,new Path (args[1]));

                        job.setMapperClass(AverageMapper.class);
job.setReducerClass(AverageReducer.class);                  job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);                 System.exit(job.waitForCompletion(true)?0:1);
            }
}

//AverageMapper.java package temperature;

import org.apache.hadoop.io.*; import org.apache.hadoop.mapreduce.*; import java.io.IOException;

public class AverageMapper extends Mapper <LongWritable, Text, Text, IntWritable>
{ public static final int MISSING = 9999;

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

//AverageReducer.java package temperature;

import org.apache.hadoop.io.IntWritable; import org.apache.hadoop.io.Text; import
org.apache.hadoop.mapreduce.*; import java.io.IOException;

public class AverageReducer extends Reducer <Text, IntWritable,Text, IntWritable>
{
        public void reduce(Text key, Iterable<IntWritable> values, Context context) throws
IOException,InterruptedException
        {
                int max_temp = 0;               int count = 0;
                for (IntWritable value : values)
                {
                        max_temp += value.get();
                        count+=1;
                }
                context.write(key, new IntWritable(max_temp/count));
        }
}
```



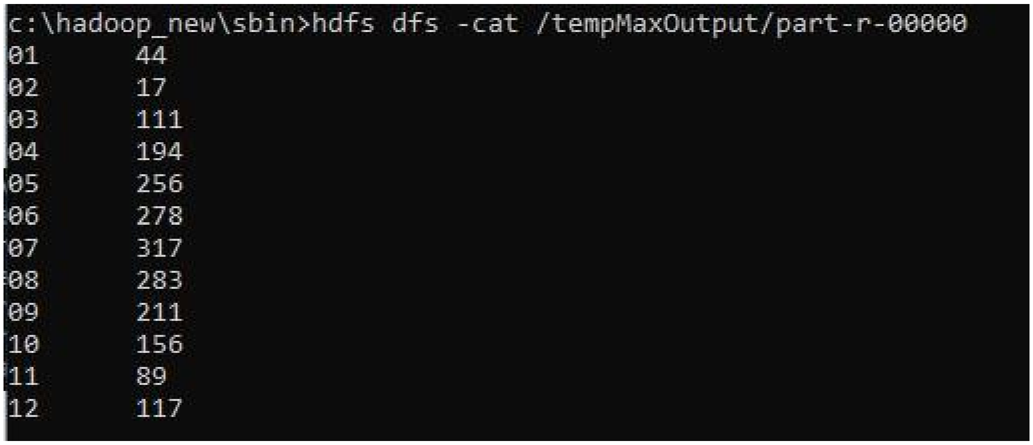
```
c:\hadoop_new\sbin>hdfs dfs -cat /tempAverageOutput/part-r-00000
1901    46
1949    94
1950    3
```

```
//TempDriver.java package
temperatureMax;


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


public class TempDriver
{       public static void main (String[] args) throws Exception
        {
                 if (args.length != 2)
```

```
                        {
                                       System.err.println("Please Enter the input and output parameters");
                                     System.exit(-1);
                        }
                        Job job = new Job();
job.setJarByClass(TempDriver.class);                  job.setJobName("Max
temperature");
                           FileInputFormat.addInputPath(job,new Path(args[0]));
                           FileOutputFormat.setOutputPath(job,new Path (args[1]));

                        job.setMapperClass(TempMapper.class);
job.setReducerClass(TempReducer.class);
                        job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
System.exit(job.waitForCompletion(true)?0:1);
            }
}

//TempMapper.java package
temperatureMax;

import org.apache.hadoop.io.*; import
org.apache.hadoop.mapreduce.*; import
java.io.IOException;

public class TempMapper extends Mapper <LongWritable, Text, Text, IntWritable>
{ public static final int MISSING = 9999;

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

```
            String line = value.toString();      String month = line.substring(19,21);
int temperature;            if (line.charAt(87)=='+')                          temperature =
Integer.parseInt(line.substring(88, 92));
            else
                        temperature = Integer.parseInt(line.substring(87, 92));    String
quality = line.substring(92, 93);  if(temperature != MISSING &&
quality.matches("[01459]"))                   context.write(new Text(month),new
IntWritable(temperature)); }

}


//TempReducer.java package
temperatureMax;


import org.apache.hadoop.io.*; import
org.apache.hadoop.mapreduce.*; import
java.io.IOException;


public class TempMapper extends Mapper <LongWritable, Text, Text, IntWritable>
{ public static final int MISSING = 9999;


public void map(LongWritable key, Text value, Context context) throws IOException,
InterruptedException
{
            String line = value.toString();      String month = line.substring(19,21);
int temperature;            if (line.charAt(87)=='+')                          temperature =
Integer.parseInt(line.substring(88, 92));
            else
                        temperature = Integer.parseInt(line.substring(87, 92));    String
quality = line.substring(92, 93);  if(temperature != MISSING &&
quality.matches("[01459]"))                   context.write(new Text(month),new
IntWritable(temperature));
```

```
        }

}
```



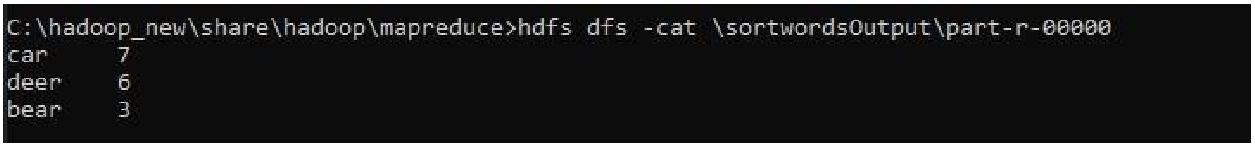
```
c:\hadoop_new\sbin>hdfs dfs -cat /tempMaxOutput/part-r-00000
01      44
02      17
03      111
04      194
05      256
06      278
07      317
08      283
09      211
10      156
11      89
12      117
```

LAB 7

For a given Text file, create a Map Reduce program to sort the content in an alphabetic order listing only top ‘n’ maximum occurrence of words.

```
// TopN.java package sortWords;

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

import java.io.IOException; import java.util.*;

public class TopN {

  public static void main(String[] args) throws Exception {
    Configuration conf = new Configuration();
    String[] otherArgs = new GenericOptionsParser(conf, args).getRemainingArgs();        if
(otherArgs.length != 2) {
      System.err.println("Usage: TopN <in> <out>");
      System.exit(2);
    }
    Job job = Job.getInstance(conf);        job.setJobName("Top N");        job.setJarByClass(TopN.class);
job.setMapperClass(TopNMapper.class);        //job.setCombinerClass(TopNReducer.class);
job.setReducerClass(TopNReducer.class);        job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
    FileInputFormat.addInputPath(job, new Path(otherArgs[0]));
    FileOutputFormat.setOutputPath(job, new Path(otherArgs[1]));
    System.exit(job.waitForCompletion(true) ? 0 : 1);
  }

  /**
   * The mapper reads one line at the time, splits it into an array of single words and emits every      *
word to the reducers with the value of 1.
   */
  public static class TopNMapper extends Mapper<Object, Text, Text, IntWritable> {

    private final static IntWritable one = new IntWritable(1);        private Text word = new Text();
    private String tokens = "[_|$#<>\\^=\\[\\]\\*/\\\\,;,.\\-:()?!\"']";

    @Override
    public void map(Object key, Text value, Context context) throws IOException,
InterruptedException {
      String cleanLine = value.toString().toLowerCase().replaceAll(tokens, " ");        StringTokenizer itr
= new StringTokenizer(cleanLine);        while (itr.hasMoreTokens()) {
        word.set(itr.nextToken().trim());        context.write(word, one);
      }
    }
  }

  /**
   * The reducer retrieves every word and puts it into a Map: if the word already exists in the    * map,
increments its value, otherwise sets it to 1.
   */
  public static class TopNReducer extends Reducer<Text, IntWritable, Text, IntWritable> {

    private Map<Text, IntWritable> countMap = new HashMap<>();

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

      // computes the number of occurrences of a single word        int sum = 0;        for (IntWritable
val : values) {        sum += val.get();
      }
      // puts the number of occurrences of this word into the map.
      // We need to create another Text object because the Text instance
      // we receive is the same for all the words        countMap.put(new Text(key), new
IntWritable(sum));
    }
@Override
    protected void cleanup(Context context) throws IOException, InterruptedException {

      Map<Text, IntWritable> sortedMap = MiscUtils.sortByValues(countMap);

      int counter = 0;        for (Text key : sortedMap.keySet()) {        if (counter++ == 3) {
break;
        }
        context.write(key, sortedMap.get(key));
      }
    }
  }

  /**
   * The combiner retrieves every word and puts it into a Map: if the word already exists in the    * map,
increments its value, otherwise sets it to 1.
   */
  public static class TopNCombiner extends Reducer<Text, IntWritable, Text, IntWritable> {

    @Override
```

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

            // computes the number of occurrences of a single word            int sum = 0;            for (IntWritable
val : values) {                sum += val.get();
            }
            context.write(key, new IntWritable(sum));
}
    }
}

// MiscUtils.java package utils;

import java.util.*;

public class MiscUtils {

    /**
sorts the map by values. Taken from:
http://javarevisited.blogspot.it/2012/12/how-to-sort-hashmap-java-by-key-and-value.html
     */
    public static <K extends Comparable, V extends Comparable> Map<K, V> sortByValues(Map<K, V>
map) {
        List<Map.Entry<K, V>> entries = new LinkedList<Map.Entry<K, V>>(map.entrySet());

        Collections.sort(entries, new Comparator<Map.Entry<K, V>>() {

            @Override            public int compare(Map.Entry<K, V> o1, Map.Entry<K, V> o2) {                return
o2.getValue().compareTo(o1.getValue());
            }
        });

        //LinkedHashMap will keep the keys in the order they are inserted
        //which is currently sorted on natural ordering
        Map<K, V> sortedMap = new LinkedHashMap<K, V>();
for (Map.Entry<K, V> entry : entries) {
            sortedMap.put(entry.getKey(), entry.getValue());
        }

        return sortedMap;
    }
}
```



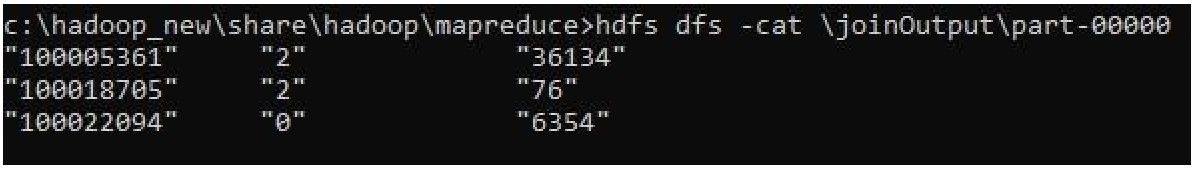
```
C:\hadoop_new\share\hadoop\mapreduce>hdfs dfs -cat \sortwordsOutput\part-r-00000
car     7
deer    6
bear    3
```

LAB 8

Create a Hadoop 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.

```
// JoinDriver.java import org.apache.hadoop.conf.Configured; import org.apache.hadoop.fs.Path; import
org.apache.hadoop.io.Text; import org.apache.hadoop.mapred.*; import
org.apache.hadoop.mapred.lib.MultipleInputs; import org.apache.hadoop.util.*;


public class JoinDriver extends Configured implements Tool {

        public static class KeyPartitioner implements Partitioner<TextPair, Text> {
                @Override
                public void configure(JobConf job) {}

                @Override
   public int getPartition(TextPair key, Text value, int numPartitions) {        return
(key.getFirst().hashCode() & Integer.MAX_VALUE) % numPartitions;
                }
        }

@Override public int run(String[] args) throws Exception {                  if (args.length != 3) {
                        System.out.println("Usage: <Department Emp Strength input>
<Department Name input> <output>");
                        return -1;
                }

                JobConf conf = new JobConf(getConf(), getClass());                  conf.setJobName("Join
'Department Emp Strength input' with 'Department Name input'");

                Path AInputPath = new Path(args[0]);
                Path BInputPath = new Path(args[1]);
                Path outputPath = new Path(args[2]);

                MultipleInputs.addInputPath(conf, AInputPath, TextInputFormat.class,
Posts.class);
                MultipleInputs.addInputPath(conf, BInputPath, TextInputFormat.class,
User.class);

                FileOutputFormat.setOutputPath(conf, outputPath);

                conf.setPartitionerClass(KeyPartitioner.class);
                conf.setOutputValueGroupingComparator(TextPair.FirstComparator.class);

                conf.setMapOutputKeyClass(TextPair.class);
        conf.setReducerClass(JoinReducer.class);

        conf.setOutputKeyClass(Text.class);

    JobClient.runJob(conf);

        return 0;
    }

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

        int exitCode = ToolRunner.run(new JoinDriver(), args);
        System.exit(exitCode);
    }
}

// JoinReducer.java import java.io.IOException; import java.util.Iterator;

import org.apache.hadoop.io.Text; import org.apache.hadoop.mapred.*;

public class JoinReducer extends MapReduceBase implements Reducer<TextPair, Text, Text, Text> {

    @Override
    public void reduce (TextPair key, Iterator<Text> values, OutputCollector<Text, Text> output,
Reporter reporter)
          throws IOException
    {

        Text nodeId = new Text(values.next());    while (values.hasNext()) {
            Text node = values.next();
        Text outValue = new Text(nodeId.toString() + "\t\t" + node.toString());
output.collect(key.getFirst(), outValue);
        }
    }
}

// User.java import java.io.IOException; import java.util.Iterator; import
org.apache.hadoop.conf.Configuration; import org.apache.hadoop.fs.FSDataInputStream; import
org.apache.hadoop.fs.FSDataOutputStream; import org.apache.hadoop.fs.FileSystem; import
org.apache.hadoop.fs.Path; import org.apache.hadoop.io.LongWritable; import
org.apache.hadoop.io.Text; import org.apache.hadoop.mapred.*;

import org.apache.hadoop.io.IntWritable;

public class User extends MapReduceBase implements Mapper<LongWritable, Text, TextPair, Text> {

    @Override
```

```
 public void map(LongWritable key, Text value, OutputCollector<TextPair, Text> output, Reporter
reporter)
                                throws IOException
          {

                    String valueString = value.toString();
                    String[] SingleNodeData = valueString.split("\t");
          output.collect(new TextPair(SingleNodeData[0], "1"), new
Text(SingleNodeData[1]));
          }
}

//Posts.java import java.io.IOException;

import org.apache.hadoop.io.*; import org.apache.hadoop.mapred.*;

public class Posts extends MapReduceBase implements Mapper<LongWritable, Text, TextPair, Text>  {

          @Override
 public void map(LongWritable key, Text value, OutputCollector<TextPair, Text> output, Reporter
reporter)
                                throws IOException
          {
                    String valueString = value.toString();
                    String[] SingleNodeData = valueString.split("\t");                    output.collect(new
TextPair(SingleNodeData[3], "0"), new
Text(SingleNodeData[9]));
          }
}

// TextPair.java import java.io.*;

import org.apache.hadoop.io.*;
public class TextPair implements WritableComparable<TextPair> {

  private Text first;   private Text second;

  public TextPair() {    set(new Text(), new Text());
  }

  public TextPair(String first, String second) {    set(new Text(first), new Text(second));
  }

  public TextPair(Text first, Text second) {    set(first, second);
  }

  public void set(Text first, Text second) {    this.first = first;    this.second = second;
  }
  public Text getFirst() {     return first;
  }

  public Text getSecond() {     return second;
  }

  @Override
  public void write(DataOutput out) throws IOException {     first.write(out);     second.write(out);
  }

  @Override   public void readFields(DataInput in) throws IOException {     first.readFields(in);
second.readFields(in);
  }

  @Override   public int hashCode() {     return first.hashCode() * 163 + second.hashCode();
  }

  @Override   public boolean equals(Object o) {     if (o instanceof TextPair) {       TextPair tp = (TextPair) o;
return first.equals(tp.first) && second.equals(tp.second);
    }     return false;
  }

  @Override   public String toString() {     return first + "\t" + second;
  }

  @Override
  public int compareTo(TextPair tp) {     int cmp = first.compareTo(tp.first);     if (cmp != 0) {       return cmp;
    }
    return second.compareTo(tp.second);
  }
  // ^^ TextPair

  // vv TextPairComparator   public static class Comparator extends WritableComparator {

    private static final Text.Comparator TEXT_COMPARATOR = new Text.Comparator();

    public Comparator() {       super(TextPair.class);
    }

    @Override     public int compare(byte[] b1, int s1, int l1,                      byte[] b2, int s2, int l2) {
          try {
      int firstL1 = WritableUtils.decodeVIntSize(b1[s1]) + readVInt(b1, s1);        int firstL2 =
WritableUtils.decodeVIntSize(b2[s2]) + readVInt(b2, s2);        int cmp = TEXT_COMPARATOR.compare(b1,
s1, firstL1, b2, s2, firstL2);        if (cmp != 0) {          return cmp;
      }
      return TEXT_COMPARATOR.compare(b1, s1 + firstL1, l1 - firstL1,
                              b2, s2 + firstL2, l2 - firstL2);
      } catch (IOException e) {         throw new IllegalArgumentException(e);
      }
    }
  }

  static {
    WritableComparator.define(TextPair.class, new Comparator());
  }
  public static class FirstComparator extends WritableComparator {

    private static final Text.Comparator TEXT_COMPARATOR = new Text.Comparator();

    public FirstComparator() {       super(TextPair.class);
    }

    @Override     public int compare(byte[] b1, int s1, int l1,                      byte[] b2, int s2, int l2) {
             try {
        int firstL1 = WritableUtils.decodeVIntSize(b1[s1]) + readVInt(b1, s1);         int firstL2 =
WritableUtils.decodeVIntSize(b2[s2]) + readVInt(b2, s2);         return TEXT_COMPARATOR.compare(b1, s1,
firstL1, b2, s2, firstL2);
      } catch (IOException e) {         throw new IllegalArgumentException(e);
      }
    }

    @Override
    public int compare(WritableComparable a, WritableComparable b) {       if (a instanceof TextPair && b
instanceof TextPair) {         return ((TextPair) a).first.compareTo(((TextPair) b).first);
      }
      return super.compare(a, b);
    }
  }
}
```



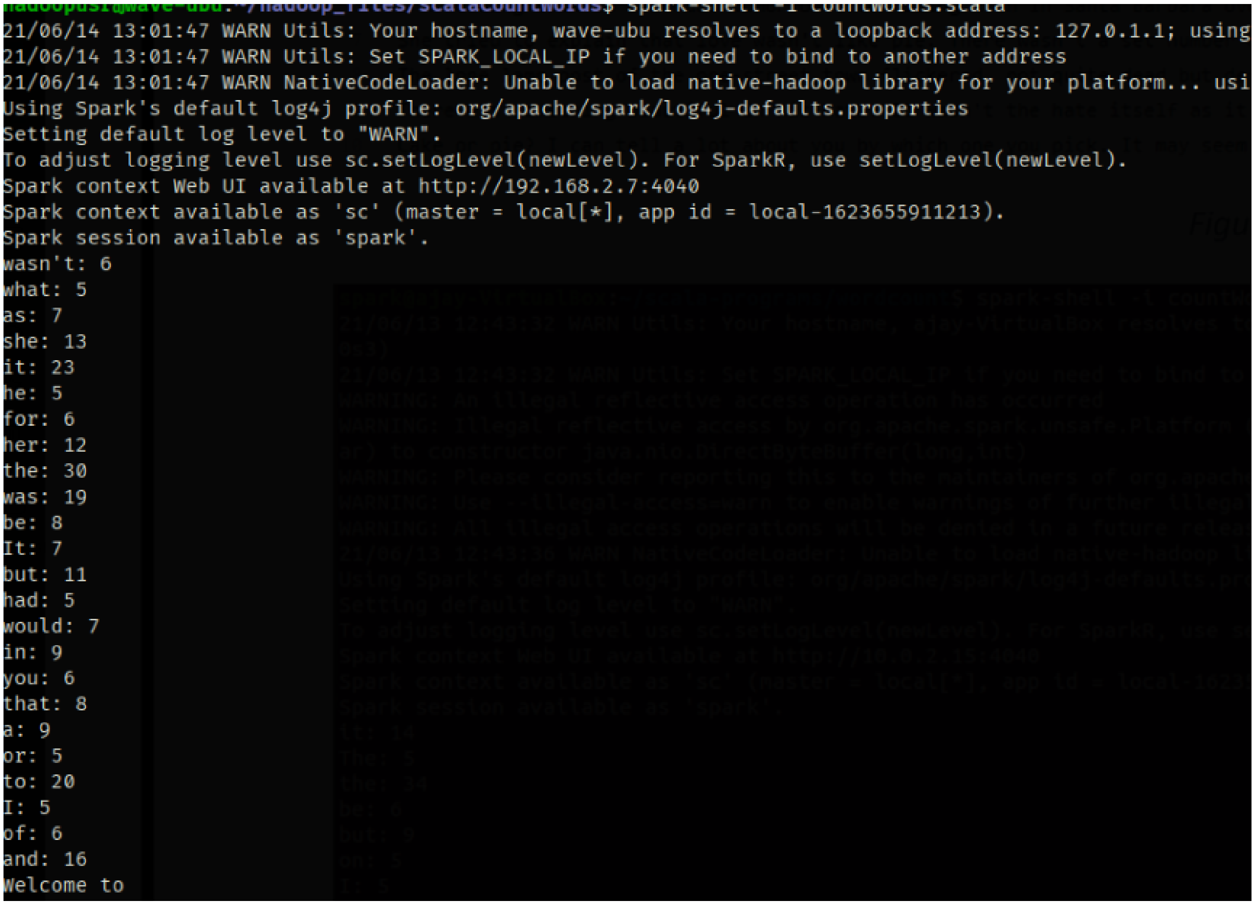
```
c:\hadoop_new\share\hadoop\mapreduce>hdfs dfs -cat \joinOutput\part-00000
"100005361"     "2"             "36134"
"100018705"     "2"             "76"
"100022094"     "0"             "6354"
```

LAB 9

Program to print word count on scala shell and print “Hello world” on scala IDE

```
scala> println("Hello World!");
Hello World!
```

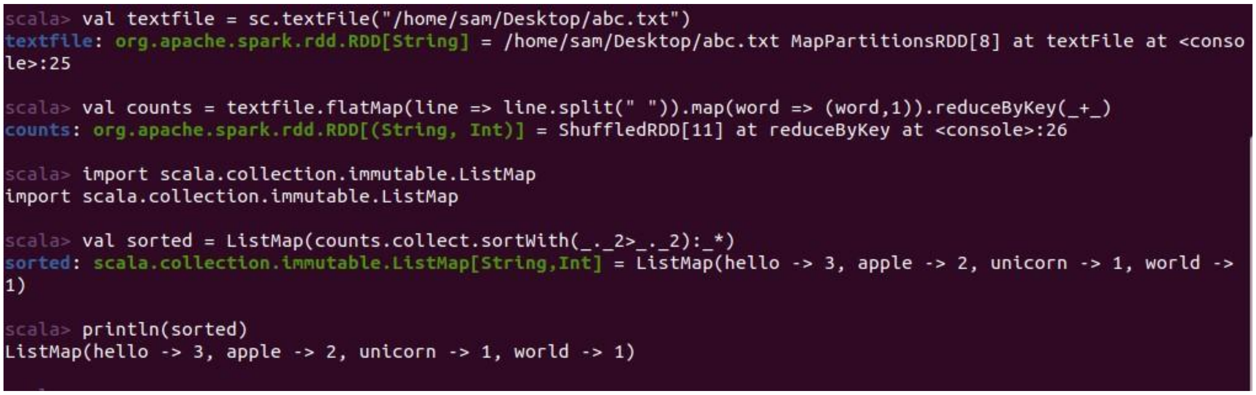
```
val data=sc.textFile("sparkdata.txt")
data.collect;
val splitdata = data.flatMap(line => line.split(" "));
splitdata.collect;
val mapdata = splitdata.map(word => (word,1));
mapdata.collect;
val reducedata = mapdata.reduceByKey(_+_);
reducedata.collect;
```



```
hadoopusr@wave-ubu:~/hadoop_files/scalacountwords$ spark-shell -i countwords.scala
21/06/14 13:01:47 WARN Utils: Your hostname, wave-ubu resolves to a loopback address: 127.0.1.1; using
21/06/14 13:01:47 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
21/06/14 13:01:47 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... usi
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://192.168.2.7:4040
Spark context available as 'sc' (master = local[*], app id = local-1623655911213).
Spark session available as 'spark'.
wasn't: 6
what: 5
as: 7
she: 13
it: 23
he: 5
for: 6
her: 12
the: 30
was: 19
be: 8
It: 7
but: 11
had: 5
would: 7
in: 9
you: 6
that: 8
a: 9
or: 5
to: 20
I: 5
of: 6
and: 16
Welcome to
```

LAB 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



```
scala> val textfile = sc.textFile("/home/sam/Desktop/abc.txt")
textfile: org.apache.spark.rdd.RDD[String] = /home/sam/Desktop/abc.txt MapPartitionsRDD[8] at textFile at <conso
le>:25

scala> val counts = textfile.flatMap(line => line.split(" ")).map(word => (word,1)).reduceByKey(_+_)
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[11] at reduceByKey at <console>:26

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(hello -> 3, apple -> 2, unicorn -> 1, world ->
1)

scala> println(sorted)
ListMap(hello -> 3, apple -> 2, unicorn -> 1, world -> 1)
```