VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELAGAVI 590018

BIG DATA ANALYTICS LAB RECORD

By

Arun D K (1BM17CS153)

Under the Guidance of

Prof. Latha NR
Assistant Professor
Department of CSE
BMS College of Engineering

Work carried out at

Department of Computer Science and Engineering
BMS College of Engineering
(Autonomous college under VTU)
P.O. Box No.: 1908, Bull Temple Road, Bangalore-560 019

2017-2018

INDEX

SL NO.	DATE	PROGRAM	PAGE NO.
1.	24-09-2020	MongoDB: Student Database	3
2.	05-10-2020	MongoDB: Customer Database	7
3.	12-10-2020	Cassandra: Employee Keyspace	11
4.	02-11-2020	Cassandra: Library Keyspace	13
5.	09-11-2020	Hadoop: Word Count	15
6.	07-12-2020	Hadoop: Average Temperature	18
7.	14-12-2020	Hive: Employee Table	20

Date: 24-09-2020

1. MongoDB: Student Database

Perform the following DB operations using MongoDB

1. **Create a database "Student" with the following attributes Rollno, Age, ContactNo, Email Id.**
2. **Insert appropriate values**
3. **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 system.**
6. **Export the created table into local file system**
7. **Drop the table**
8. **Import a given csv dataset from local file system into mongodb collection.**

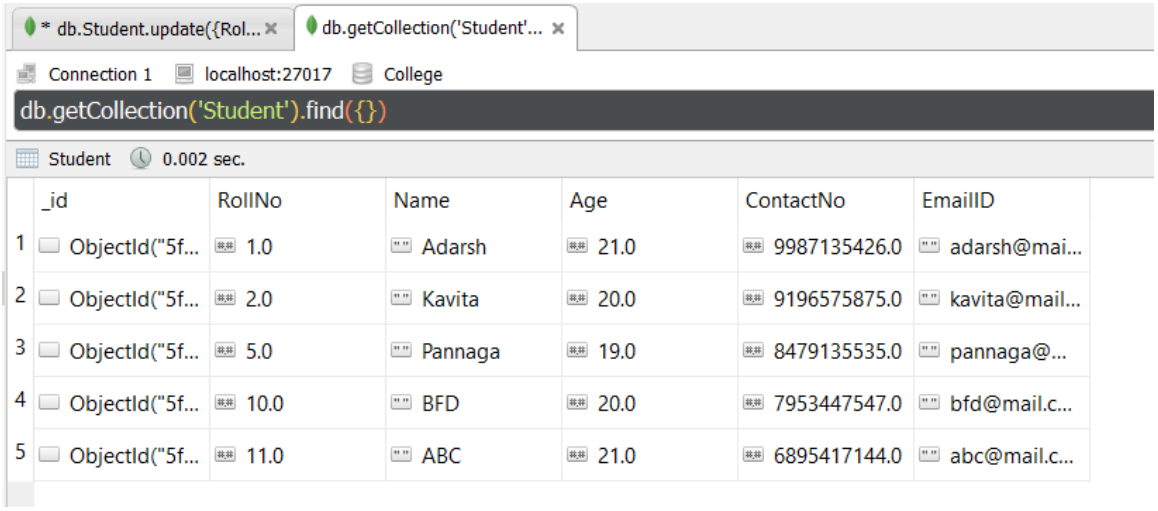
```
use StudentDB
```

1. Create a database "Student" with the following attributes Rollno, Age, ContactNo, Email-Id

```
db.createCollection("Student")
```

2. Insert appropriate values

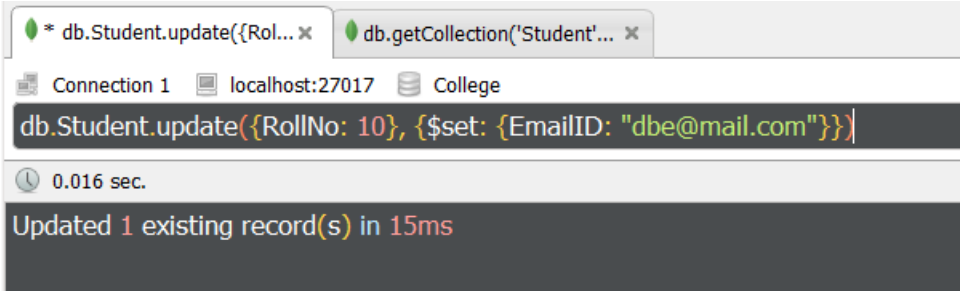
```
db.Student.insertMany([{RollNo: 1, Name: "Adarsh", Age: 21, ContactNo:
9987135426, EmailID: "adarsh@mail.com"},
{RollNo: 2, Name: "Kavita", Age: 20, ContactNo: 9196575875, EmailID:
"kavita@mail.com"},
{RollNo: 5, Name: "Pannaga", Age: 19, ContactNo: 8479135535, EmailID:
"pannaga@mail.com"},
{RollNo: 10, Name: "BFD", Age: 20, ContactNo: 7953447547, EmailID:
"bfd@mail.com"},
{RollNo: 11, Name: "ABC", Age: 21, ContactNo: 6895417144, EmailID:
"abc@mail.com"}])
db.Student.find()
```



	_id	RollNo	Name	Age	ContactNo	EmailID
1	ObjectId("5f...	1.0	Adarsh	21.0	9987135426.0	adarsh@mai...
2	ObjectId("5f...	2.0	Kavita	20.0	9196575875.0	kavita@mail...
3	ObjectId("5f...	5.0	Pannaga	19.0	8479135535.0	pannaga@...
4	ObjectId("5f...	10.0	BFD	20.0	7953447547.0	bfd@mail.c...
5	ObjectId("5f...	11.0	ABC	21.0	6895417144.0	abc@mail.c...

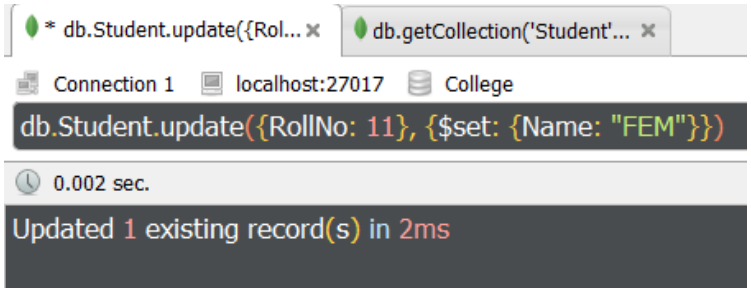
3. Write query to update Email-Id of a student with rollno 10

```
db.Student.update({RollNo:10}, {$set:{EmailID:“dbe@gmail.com”}})
db.Student.find({RollNo:10})
```



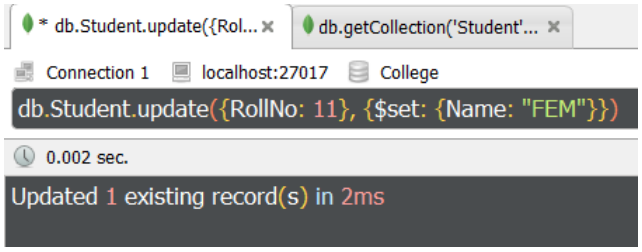
4. Replace the student name from “ABC” to “FEM” of rollno 11.

```
db.Student.update({RollNo:11},{$set:{Name:"FEM"}});
db.Student.find({RollNo:11})
```



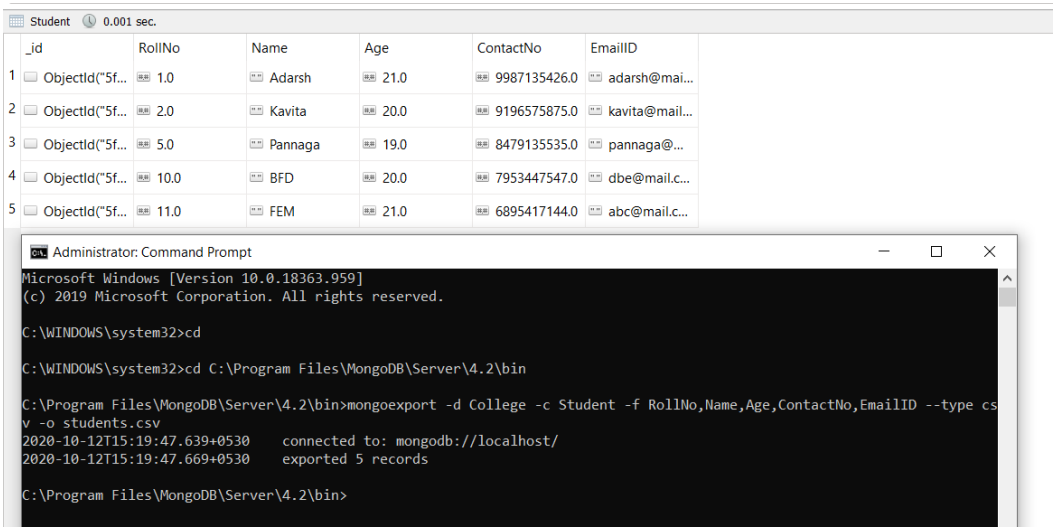
5. Replace the student name from “ABC” to “FEM” of rollno 11.

```
db.Student.update({RollNo:11},{$set:{Name:"FEM"}});
db.Student.find({RollNo:11})
```



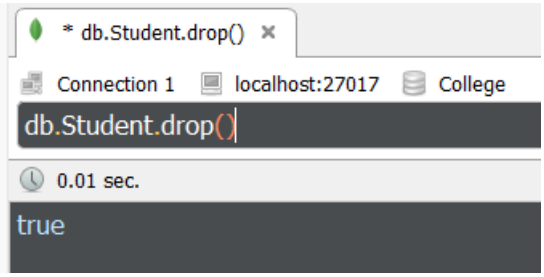
6. Export the created table into local file system

```
mongoexport - d College - c Student - f RollNo, Name, Age, ContactNo,
EmailID--type csv - o students.csv
```



7. Drop the table

```
db.Student.drop()
```



8. Import a given csv dataset from local file system into mongodb collection

```
mongoimport - d College - c Student--type csv--file students.csv--
headerline
```

Connection 1 localhost:27017 College

```
db.getCollection('Student').find({})
```

Student 0.002 sec.

	_id	RollNo	Name	Age	ContactNo	EmailID
1	ObjectId("5f...	2	Kavita	20	9196575875.0	kavita@mail...
2	ObjectId("5f...	5	Pannaga	19	8479135535.0	pannaga@...
3	ObjectId("5f...	10	BFD	20	7953447547.0	dbe@mail.c...
4	ObjectId("5f...	11	FEM	21	6895417144.0	abc@mail.c...
5	ObjectId("5f...	1	Adarsh	21	9987135426.0	adarsh@mai...

Administrator: Command Prompt

```
Microsoft Windows [Version 10.0.18363.959]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>cd

C:\WINDOWS\system32>cd C:\Program Files\MongoDB\Server\4.2\bin

C:\Program Files\MongoDB\Server\4.2\bin>mongoexport -d College -c Student -f RollNo,Name,Age,ContactNo,EmailID --type cs
v -o students.csv
2020-10-12T15:19:47.639+0530    connected to: mongodb://localhost/
2020-10-12T15:19:47.669+0530    exported 5 records

C:\Program Files\MongoDB\Server\4.2\bin>mongoimport -d College -c Student --type csv --file students.csv --headerline
2020-10-12T15:24:24.478+0530    connected to: mongodb://localhost/
2020-10-12T15:24:24.522+0530    5 document(s) imported successfully. 0 document(s) failed to import.

C:\Program Files\MongoDB\Server\4.2\bin>
```

Date: 05-10-2020

2. MongoDB: Customer Database

Perform the following DB operations using MongoDB.

1. **Create a collection by name Customers with the following attributes. Cust_id, Acc_Bal, Acc_Type**
2. **Insert at least 5 values into the table**
3. **Write a query to display those records whose total account balance is greater than 1200 of account type 'Z' for each customer_id.**
4. **Determine Minimum and Maximum account balance for each customer_id.**
5. **Export the created collection into local file system**
6. **Drop the table**
7. **Import a given csv dataset from local file system into mongodb collection.**

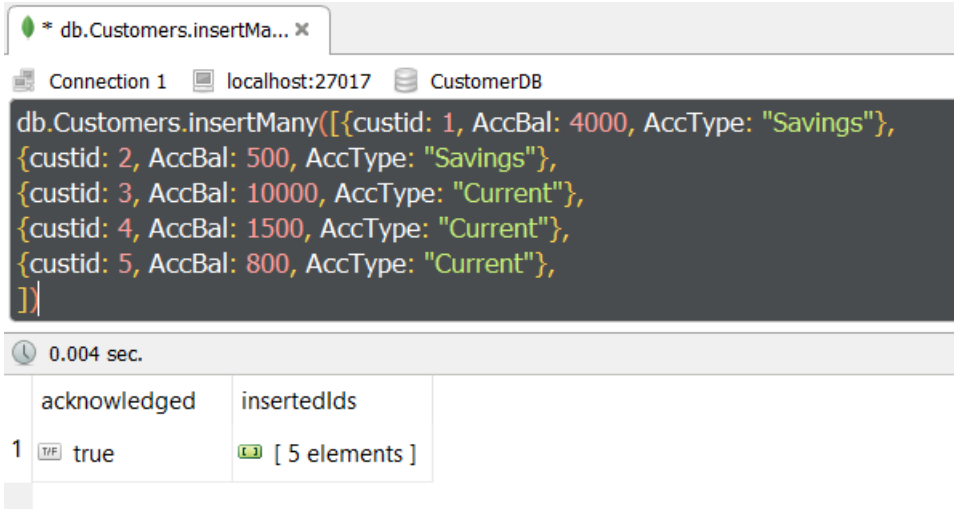
```
use CustomerDB
```

1. Create a collection by name Customers with the following attributes.Cust_id, Acc_Bal, Acc_Type

```
db.createCollection("Customers")
```

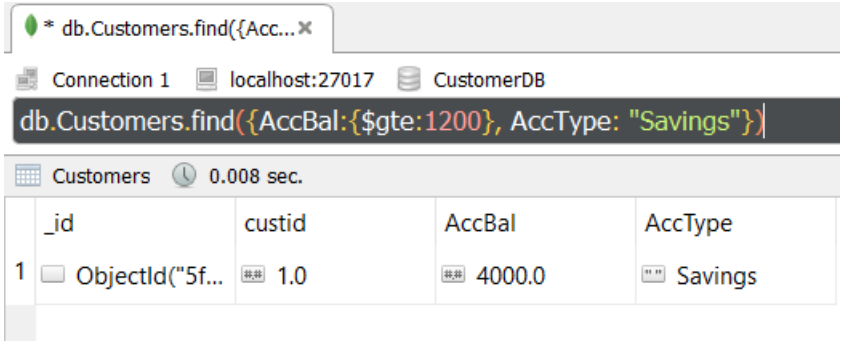
2. Insert at least 5 values into the table

```
db.Customers.insertMany([{custid: 1, AccBal: 4000, AccType: "Savings"},
{custid: 2, AccBal: 500, AccType: "Savings"},
{custid: 3, AccBal: 10000, AccType: "Current"},
{custid: 4, AccBal: 1500, AccType: "Current"},
{custid: 5, AccBal: 800, AccType: "Current"}])
db.Customers.find()
```



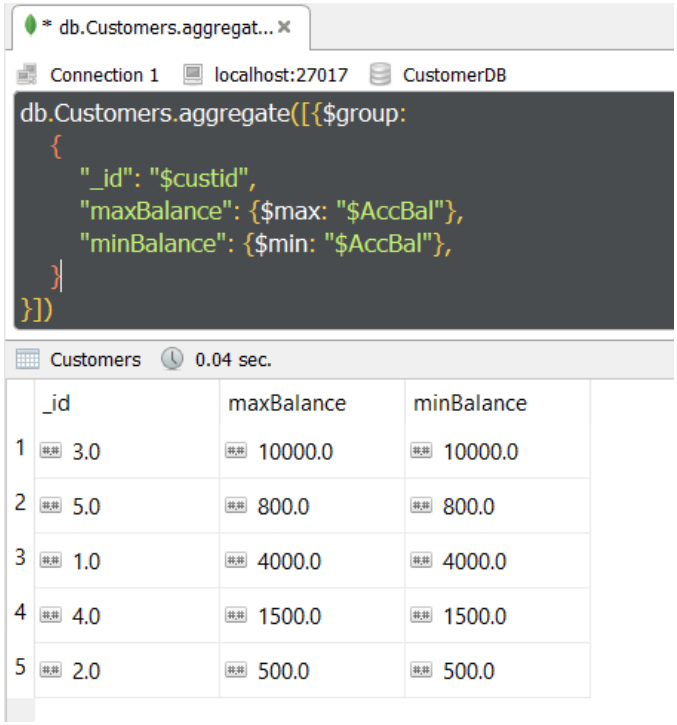
3. Write a query to display those records whose total account balance is greater than 1200 of account type ‘Z’ for each customer_id.

```
db.Customers.find({AccBal:{$gte:1200}, AccType:"Savings"})
```



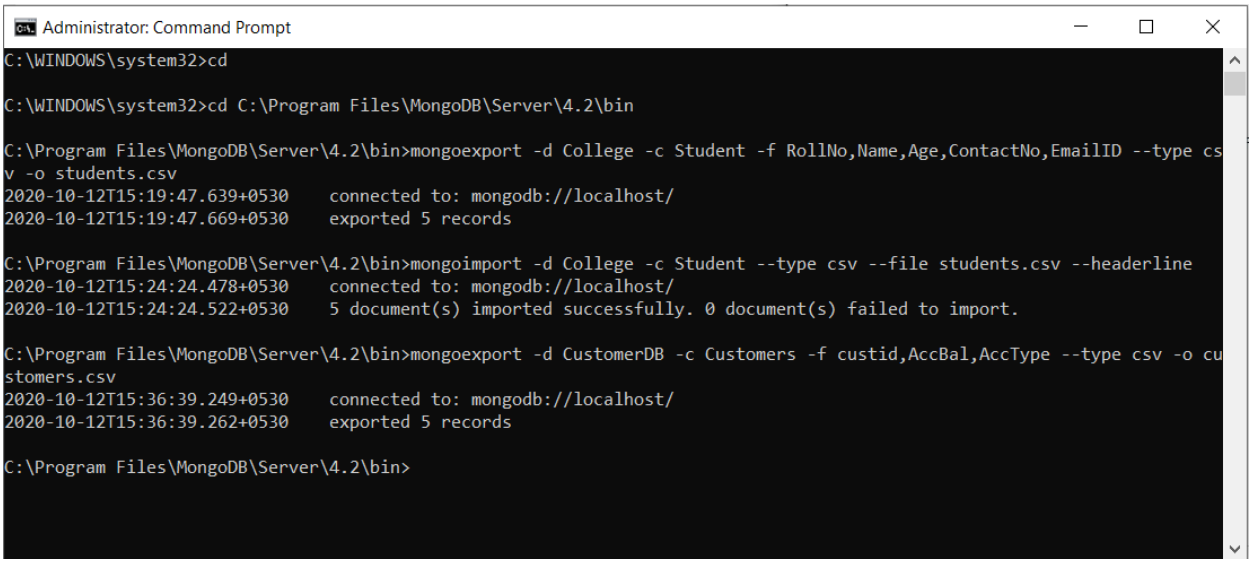
4. Determine Minimum and Maximum account balance for each customer_id.

```
db.Customers.aggregate([{
    $group:
        {_id: "$custid",
        maxBalance: {$max: "$AccBal"},
        minBalance: {$min: "$AccBal"}}
    }]);
```



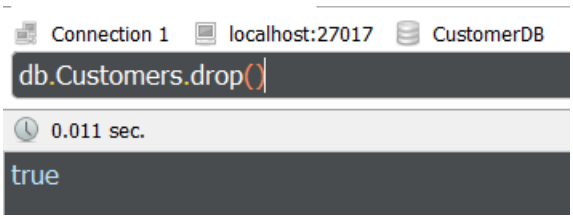
5. Export the created collection into local file system

```
mongoexport -d CustomerDB -c Customers -f custid,AccBal,AccType --type
csv -o customers.csv
```



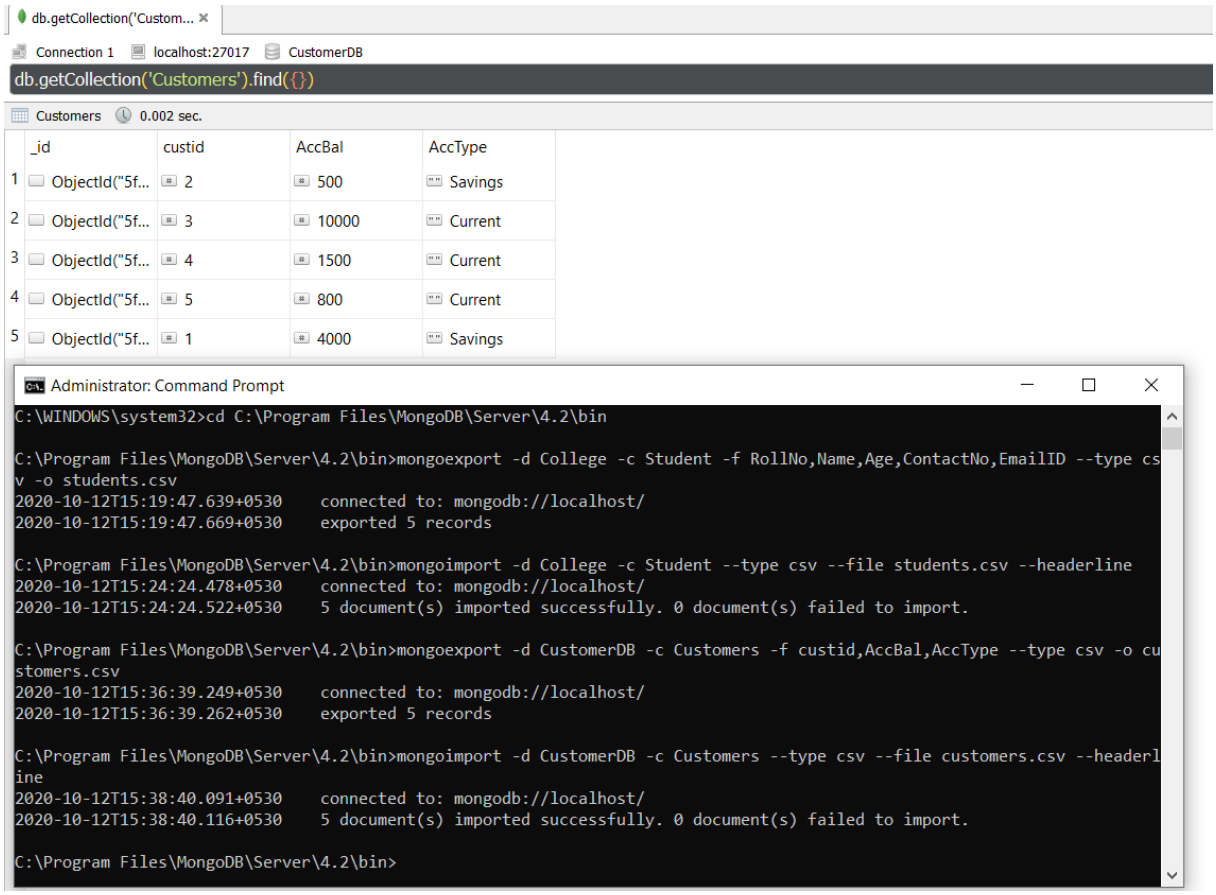
6. Drop the table

```
db.Customers.drop()
```



7. Import a given csv dataset from local file system into mongodb collection

```
mongoimport -d CustomerDB -c Customers --type csv --file customers.csv --
headerline
```



Date: 12-10-2020

3. Cassandra: Employee Keyspace

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 3. Update Employee name and Department of Emp-Id 121**
4. **Sort the details of Employee records based on salary**
5. **. Alter the schema of the table Employee_Info to add a column Projects which stores a set of Projects done by the corresponding Employee.**
6. **Update the altered table to add project names.**
7. **Create a TTL of 15 seconds to display the values of Employees.**

1. Create a keyspace by name Employee

```
Administrator: Command Prompt - cqlsh
Microsoft Windows [Version 10.0.18363.1139]
(c) 2019 Microsoft Corporation. All rights reserved.

C:\WINDOWS\system32>cd C:\apache-cassandra-3.11.4\bin

C:\apache-cassandra-3.11.4\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 3.11.4 | CQL spec 3.4.4 | Native protocol v4]
Use HELP for help.
cqlsh> CREATE KEYSPACE Employees WITH replication ={'class':'SimpleStrategy','replication_factor':3};
cqlsh> use Employees;
```

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 Employee_Info (Emp_Id int PRIMARY KEY, Emp_Name text,Designation text, DateOfJoining timest
amp, Salary double,Dept_Name text);
cqlsh:employee> DESCRIBE TABLES;

employee_info
```

3. Insert the values into the table in batch

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

5. Alter the schema of the table Employee_Info to add a column Projects which stores a set of Projects done by the corresponding Employee.

```
Command Prompt - cqlsh
cqlsh:employee> select * from employee_info;

 emp_id | dept_name | designation | doj | emp_name | salary
--------+-----------+-------------+-----+----------+--------

(0 rows)
cqlsh:employee> BEGIN BATCH
            ... INSERT INTO employee_info(emp_id,dept_name,designation,doj,emp_name,salary)values(120,'Development','CTO','10/09/2015','Piyush',2000000);
            ... INSERT INTO employee_info(emp_id,dept_name,designation,doj,emp_name,salary)values(121,'HR','Employee','15/09/2012','Akash',1500000);
            ... INSERT INTO employee_info(emp_id,dept_name,designation,doj,emp_name,salary)values(122,'Maintainance staff','03/08/2016','Chinmayi',53000);
            ... INSERT INTO employee_info(emp_id,dept_name,designation,doj,emp_name,salary)values(123,'IT','Assistant','24/09/2019','Trisha',100000);
            ... APPLY BATCH;
InvalidRequest: Error from server: code=2200 [Invalid query] message="Unmatched column names/values"
cqlsh:employee> BEGIN BATCH
            ... INSERT INTO employee_info(emp_id,dept_name,designation,doj,emp_name,salary)values(120,'Development','CTO','10/09/2015','Piyush',2000000);
            ... APPLY BATCH;
cqlsh:employee> BEGIN BATCH
            ... INSERT INTO employee_info(emp_id,dept_name,designation,doj,emp_name,salary)values(121,'HR','Employee','15/09/2012','Akash',1500000);
            ... INSERT INTO employee_info(emp_id,dept_name,designation,doj,emp_name,salary)values(123,'IT','Assistant','24/09/2019','Trisha',100000);
            ... APPLY BATCH;
cqlsh:employee> INSERT INTO employee_info(emp_id,dept_name,designation,doj,emp_name,salary)values(122,'Maintainance staff','03/08/2016','Chinmayi',53000);
InvalidRequest: Error from server: code=2200 [Invalid query] message="Unmatched column names/values"
cqlsh:employee> INSERT INTO employee_info(emp_id,dept_name,designation,doj,emp_name,salary)values(122,'MaintainanceStaff','03/08/2016','Chinmayi',53000);
InvalidRequest: Error from server: code=2200 [Invalid query] message="Unmatched column names/values"
cqlsh:employee> INSERT INTO employee_info(emp_id,dept_name,designation,doj,emp_name,salary)values(122,'Maintainance', 'Staff','03/08/2016','Chinmayi',53000);
cqlsh:employee> select * from employee_info;

 emp_id | dept_name    | designation | doj        | emp_name | salary
--------+--------------+-------------+------------+----------+--------
    120 |  Development |         CTO | 10/09/2015 |   Piyush | 2000000
    123 |           IT |   Assistant | 24/09/2019 |   Trisha |  100000
    122 | Maintainance |       Staff | 03/08/2016 | Chinmayi |   53000
    121 |           HR |    Employee | 15/09/2012 |    Akash | 1500000

(4 rows)
cqlsh:employee> UPDATE employee_info SET emp_name = 'Aakash',dept_name = 'IT' WHERE emp_id = 121;
cqlsh:employee> select * from employee_info;

 emp_id | dept_name    | designation | doj        | emp_name | salary
--------+--------------+-------------+------------+----------+--------
    120 |  Development |         CTO | 10/09/2015 |   Piyush | 2000000
    123 |           IT |   Assistant | 24/09/2019 |   Trisha |  100000
    122 | Maintainance |       Staff | 03/08/2016 | Chinmayi |   53000
    121 |           IT |    Employee | 15/09/2012 |   Aakash | 1500000

(4 rows)
cqlsh:employee> ALTER TABLE employee_info ADD Project VARCHAR;
```

6. Update the altered table to add project names.

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

```
Command Prompt - cqlsh
cqlsh:employee> UPDATE employee_info SET project='Facial recognition' WHERE emp_id=123;
cqlsh:employee> UPDATE employee_info SET project='Blockchain' WHERE emp_id=122;
cqlsh:employee> select * from employee_info;

 emp_id | dept_name    | designation | doj        | emp_name | project            | salary
--------+--------------+-------------+------------+----------+--------------------+--------
    120 |  Development |         CTO | 10/09/2015 |   Piyush |                TIP | 2000000
    123 |           IT |   Assistant | 24/09/2019 |   Trisha | Facial recognition |  100000
    122 | Maintainance |       Staff | 03/08/2016 | Chinmayi |         Blockchain |   53000
    121 |           IT |    Employee | 15/09/2012 |   Aakash | Sentiment Analysis | 1500000

(4 rows)
cqlsh:employee> INSERT INTO employee_info(emp_id, dept_name, designation, doj, emp_name, project,salary)values(124, 'PR','Senior Manager','01/05/2010','Load Balancing Server','Pranav',500000) USING TTL 60;
cqlsh:employee> SELECT TTL(desgination) FROM employee_info where emp_id=124;
InvalidRequest: Error from server: code=2200 [Invalid query] message="Undefined column name desgination"
cqlsh:employee> SELECT TTL(designation) FROM employee_info where emp_id=124;

 ttl(designation)
------------------
               37

(1 rows)
cqlsh:employee> SELECT TTL(designation) FROM employee_info where emp_id=124;

 ttl(designation)
------------------
               29

(1 rows)
cqlsh:employee> select * from employee_info;

 emp_id | dept_name    | designation    | doj        | emp_name              | project            | salary
--------+--------------+----------------+------------+-----------------------+--------------------+--------
    120 |  Development |            CTO | 10/09/2015 |                Piyush |                TIP | 2000000
    123 |           IT |      Assistant | 24/09/2019 |                Trisha | Facial recognition |  100000
    122 | Maintainance |          Staff | 03/08/2016 |               Chinmayi |         Blockchain |   53000
    121 |           IT |       Employee | 15/09/2012 |                Aakash | Sentiment Analysis | 1500000
    124 |           PR | Senior Manager | 01/05/2010 | Load Balancing Server |             Pranav |  500000

(5 rows)
cqlsh:employee> select * from employee_info;

 emp_id | dept_name    | designation | doj        | emp_name | project            | salary
--------+--------------+-------------+------------+----------+--------------------+--------
    120 |  Development |         CTO | 10/09/2015 |   Piyush |                TIP | 2000000
    123 |           IT |   Assistant | 24/09/2019 |   Trisha | Facial recognition |  100000
    122 | Maintainance |       Staff | 03/08/2016 | Chinmayi |         Blockchain |   53000
    121 |           IT |    Employee | 15/09/2012 |   Aakash | Sentiment Analysis | 1500000

(4 rows)
```

Date: 02-11-2020

4. Cassandra: Library Keyspace

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**

Command Prompt - cqlsh

```
Connected to Test Cluster at 127.0.0.1:9042.
[cqlsh 5.0.1 | Cassandra 3.11.8 | CQL spec 3.4.4 | Native protocol v4]
Use HELP for help.
WARNING: pyreadline dependency missing.  Install to enable tab completion.
cqlsh> CREATE KEYSPACE Library WITH replication = {'class':'SimpleStrategy','replication_factor':3};
cqlsh> CREATE COLUMNFAMILY libraryinfo(stud_id uuid, counter_value counter, stud_name VARCHAR, book_name VARCHAR, book_id INT, doi VARCHAR, PRIMARY KEY(stud_id,stud_name,book_name,book_id,doi));
InvalidRequest: Error from server: code=2200 [Invalid query] message="No keyspace has been specified. USE a keyspace, or explicitly specify keyspace.tablename"
cqlsh> USE Library;
cqlsh:library> CREATE COLUMNFAMILY libraryinfo(stud_id uuid, counter_value counter, stud_name VARCHAR, book_name VARCHAR, book_id INT, doi VARCHAR, PRIMARY KEY(stud_id,stud_name,book_name,book_id,doi));
cqlsh:library>
Traceback (most recent call last):
  File "C:\Program Files\apache-cassandra-3.11.8\bin\\cqlsh.py", line 2458, in <module>
    main(*read_options(sys.argv[1:], os.environ))
  File "C:\Program Files\apache-cassandra-3.11.8\bin\\cqlsh.py", line 2446, in main
    shell.cmdloop()
  File "C:\Program Files\apache-cassandra-3.11.8\bin\\cqlsh.py", line 893, in cmdloop
    self.handle_eof()
  File "C:\Program Files\apache-cassandra-3.11.8\bin\\cqlsh.py", line 939, in handle_eof
    statement = self.statement.getvalue()
  File "C:\Python27\lib\StringIO.py", line 269, in getvalue
    _complain_ifclosed(self.closed)
  File "C:\Python27\lib\StringIO.py", line 38, in _complain_ifclosed
    def _complain_ifclosed(closed):
KeyboardInterrupt
Terminate batch job (Y/N)? n
C:\Program Files\apache-cassandra-3.11.8\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 3.11.8 | CQL spec 3.4.4 | Native protocol v4]
Use HELP for help.
WARNING: pyreadline dependency missing.  Install to enable tab completion.
cqlsh> USE Library;
cqlsh:library> UPDATE library_info set counter_value=counter_value+1 where stud_id=uuid() and stud_name='Anjali' and book_name='Inferno' and book_id=1 and doi='20/09/2020';
InvalidRequest: Error from server: code=2200 [Invalid query] message="unconfigured table library_info"
cqlsh:library> CREATE COLUMNFAMILY libraryinfo(stud_id uuid, counter_value counter, stud_name VARCHAR, book_name VARCHAR, book_id INT, doi VARCHAR, PRIMARY KEY(stud_id,stud_name,book_name,book_id,doi));
AlreadyExists: Table 'library.libraryinfo' already exists
cqlsh:library> UPDATE library_info set counter_value=counter_value+1 where stud_id=uuid() and stud_name='Anjali' and book_name='Inferno' and book_id=1 and doi='20/09/2020';
InvalidRequest: Error from server: code=2200 [Invalid query] message="unconfigured table library_info"
cqlsh:library> UPDATE libraryinfo set counter_value=counter_value+1 where stud_id=uuid() and stud_name='Anjali' and book_name='Inferno' and book_id=1 and doi='20/09/2020';
cqlsh:library> UPDATE libraryinfo set counter_value=counter_value+1 where stud_id=uuid() and stud_name='Ashwini' and book_name='BDA' and book_id=2 and doi='25/09/2020';
cqlsh:library> UPDATE libraryinfo set counter_value=counter_value+1 where stud_id=uuid() and stud_name='Atharv' and book_name='The Alchemist' and book_id=3 and doi='28/09/2020';
cqlsh:library> select * from libraryinfo;
```

Command Prompt - cqlsh

```
 stud_id                              | stud_name | book_name     | book_id | doi        | counter_value
--------------------------------------+-----------+---------------+---------+------------+---------------
 03408842-8701-4e1f-ac4e-81071d962393 |    Anjali |       Inferno |       1 | 20/09/2020 |             1
 f588032e-98da-45c8-95c1-6e30ccd3c2bb |    Atharv | The Alchemist |       3 | 28/09/2020 |             1
 515130bd-6a01-4d3c-824f-e671f6d4c553 |   Ashwini |           BDA |       2 | 25/09/2020 |             1

(3 rows)
cqlsh:library> UPDATE libraryinfo set counter_value=counter_value+1 where stud_id=515130bd-6a01-4d3c-824f-e671f6d4c553 and stud_name='Ashwini' and book_name='BDA' and book_id=2 and doi='25/09/2020';
cqlsh:library> select * from libraryinfo where counter_value = 2 ALLOW FILTERING;

 stud_id                              | stud_name | book_name | book_id | doi        | counter_value
--------------------------------------+-----------+-----------+---------+------------+---------------
 515130bd-6a01-4d3c-824f-e671f6d4c553 |   Ashwini |       BDA |       2 | 25/09/2020 |             2

(1 rows)
cqlsh:library> COPY libraryinfo(stud_id,counter_value,stud_name,book_name,book_id,doi) TO 'C:\Users\arund\Documents\res.csv' WITH HEADER =TRUE;
Using 3 child processes

Starting copy of library.libraryinfo with columns [stud_id, counter_value, stud_name, book_name, book_id, doi].
Processed: 3 rows; Rate:       6 rows/s; Avg. rate:       3 rows/s
3 rows exported to 1 files in 0.998 seconds.
```

Command Prompt - cqlsh

```
cqlsh:library> COPY libraryinfo(stud_id,counter_value,stud_name,book_name,book_id,doi) FROM 'C:\Users\arund\Documents\res.csv' WITH HEADER =TRUE;
Using 3 child processes

Starting copy of library.libraryinfo with columns [stud_id, counter_value, stud_name, book_name, book_id, doi].
Process ImportProcess-6:       4 rows/s; Avg. rate:       4 rows/s
PProcess ImportProcess-5:
Trocess ImportProcess-4:
Traceback (most recent call last):
T raceback (most recent call last):
raceback (most recent call last):
 File "C:\Python27\lib\multiprocessing\process.py", line 267, in _bootstrap
   File "C:\Python27\lib\multiprocessing\process.py", line 267, in _bootstrap
 File "C:\Python27\lib\multiprocessing\process.py", line 267, in _bootstrap
      self.run()
   self.run()
   self.run()
   File "C:\Program Files\apache-cassandra-3.11.8\bin\..\pylib\cqlshlib\copyutil.py", line 2328, in run
   File "C:\Program Files\apache-cassandra-3.11.8\bin\..\pylib\cqlshlib\copyutil.py", line 2328, in run
 File "C:\Program Files\apache-cassandra-3.11.8\bin\..\pylib\cqlshlib\copyutil.py", line 2328, in run
   self.close()
      self.close()
 File "C:\Program Files\apache-cassandra-3.11.8\bin\..\pylib\cqlshlib\copyutil.py", line 2332, in close
   self.close()
    File "C:\Program Files\apache-cassandra-3.11.8\bin\..\pylib\cqlshlib\copyutil.py", line 2332, in close
   self._session.cluster.shutdown()
 File "C:\Program Files\apache-cassandra-3.11.8\bin\..\pylib\cqlshlib\copyutil.py", line 2332, in close
    File "C:\Program Files\apache-cassandra-3.11.8\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\cluster.py", line 1259, in shutdown
   self._session.cluster.shutdown()
   self._session.cluster.shutdown()
      self.control_connection.shutdown()
 File "C:\Program Files\apache-cassandra-3.11.8\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:\Program Files\apache-cassandra-3.11.8\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:\Program Files\apache-cassandra-3.11.8\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\cluster.py", line 2850, in shutdown
   self.control_connection.shutdown()
      self._connection.close()
 File "C:\Program Files\apache-cassandra-3.11.8\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\cluster.py", line 2850, in shutdown
   self.control_connection.shutdown()
    File "C:\Program Files\apache-cassandra-3.11.8\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:\Program Files\apache-cassandra-3.11.8\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\cluster.py", line 2850, in shutdown
     AsyncoreConnection.create_timer(0, partial(asyncore.dispatcher.close, self))
   self._connection.close()
     self._connection.close()
 File "C:\Program Files\apache-cassandra-3.11.8\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:\Program Files\apache-cassandra-3.11.8\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:\Program Files\apache-cassandra-3.11.8\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
     AsyncoreConnection.create_timer(0, partial(asyncore.dispatcher.close, self))
     AsyncoreConnection.create_timer(0, partial(asyncore.dispatcher.close, self))
   cls._loop.add_timer(timer)
 File "C:\Program Files\apache-cassandra-3.11.8\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
 A File "C:\Program Files\apache-cassandra-3.11.8\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
```

Command Prompt - cqlsh

```
raceback (most recent call last):
 File "C:\Python27\lib\multiprocessing\process.py", line 267, in _bootstrap
   File "C:\Python27\lib\multiprocessing\process.py", line 267, in _bootstrap
 File "C:\Python27\lib\multiprocessing\process.py", line 267, in _bootstrap
      self.run()
   self.run()
   self.run()
   File "C:\Program Files\apache-cassandra-3.11.8\bin\..\pylib\cqlshlib\copyutil.py", line 2328, in run
   File "C:\Program Files\apache-cassandra-3.11.8\bin\..\pylib\cqlshlib\copyutil.py", line 2328, in run
 File "C:\Program Files\apache-cassandra-3.11.8\bin\..\pylib\cqlshlib\copyutil.py", line 2328, in run
   self.close()
      self.close()
 File "C:\Program Files\apache-cassandra-3.11.8\bin\..\pylib\cqlshlib\copyutil.py", line 2332, in close
   self.close()
    File "C:\Program Files\apache-cassandra-3.11.8\bin\..\pylib\cqlshlib\copyutil.py", line 2332, in close
   self._session.cluster.shutdown()
 File "C:\Program Files\apache-cassandra-3.11.8\bin\..\pylib\cqlshlib\copyutil.py", line 2332, in close
    File "C:\Program Files\apache-cassandra-3.11.8\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\cluster.py", line 1259, in shutdown
   self._session.cluster.shutdown()
   self._session.cluster.shutdown()
      self.control_connection.shutdown()
 File "C:\Program Files\apache-cassandra-3.11.8\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:\Program Files\apache-cassandra-3.11.8\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:\Program Files\apache-cassandra-3.11.8\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\cluster.py", line 2850, in shutdown
   self.control_connection.shutdown()
      self._connection.close()
 File "C:\Program Files\apache-cassandra-3.11.8\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\cluster.py", line 2850, in shutdown
   self.control_connection.shutdown()
    File "C:\Program Files\apache-cassandra-3.11.8\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:\Program Files\apache-cassandra-3.11.8\bin\..\lib\cassandra-driver-internal-only-3.11.0-bb96859b.zip\cassandra-driver-3.11.0-bb96859b\cassandra\cluster.py", line 2850, in shutdown
     AsyncoreConnection.create_timer(0, partial(asyncore.dispatcher.close, self))
   self._connection.close()
     self._connection.close()
 File "C:\Program Files\apache-cassandra-3.11.8\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:\Program Files\apache-cassandra-3.11.8\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:\Program Files\apache-cassandra-3.11.8\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
     AsyncoreConnection.create_timer(0, partial(asyncore.dispatcher.close, self))
     AsyncoreConnection.create_timer(0, partial(asyncore.dispatcher.close, self))
   cls._loop.add_timer(timer)
 File "C:\Program Files\apache-cassandra-3.11.8\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
 A File "C:\Program Files\apache-cassandra-3.11.8\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'
    cls._loop.add_timer(timer)
   cls._loop.add_timer(timer)
AAttributeError: 'NoneType' object has no attribute 'add_timer'
ttributeError: 'NoneType' object has no attribute 'add_timer'
Processed: 3 rows; Rate:       2 rows/s; Avg. rate:       3 rows/s
3 rows imported from 1 files in 1.058 seconds (0 skipped).
cqlsh:library>
```

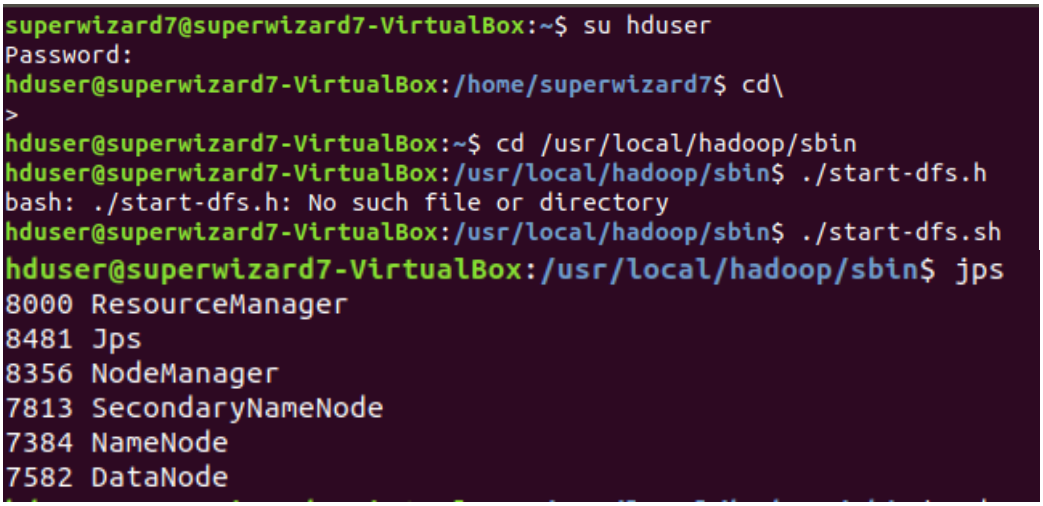
Date: 09-11-2020

5. Hadoop: Word Count

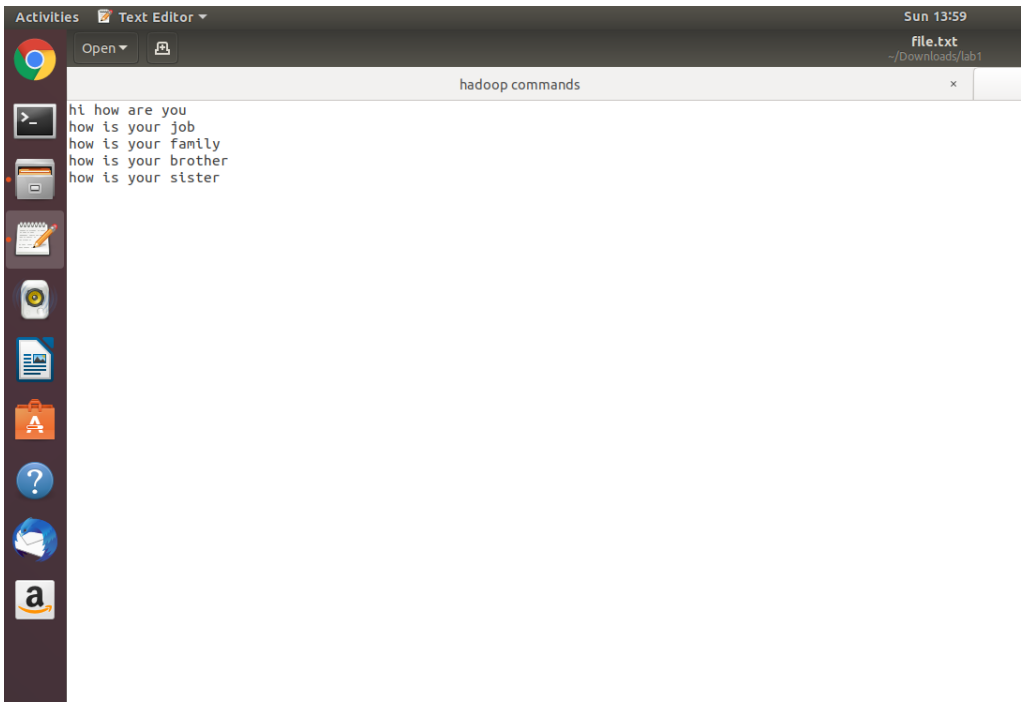
Hadoop program to find the word count

1. Starting Hadoop Cluster

```
$ su hduser
$ cd\
$ start-all.sh
```



2. Creating a file to count words



3. Moving file to Hadoop system

```
$ hadoop fs –mkdir /lab1
$ hadoop fs –ls /
$ hadoop fs -copyFromLocal /home/superwizard7/Downloads/lab1/file.txt
/lab1/input.txt
```

```
hduser@superwizard7-VirtualBox:~$ hdfs dfs -mkdir /lab1
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 reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/12/14 15:33:41 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

```
hduser@superwizard7-VirtualBox:~$ hdfs dfs -copyFromLocal /home/superwizard7/Downloads/lab1/file.txt /lab1/input.txt
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 reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/12/14 15:36:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

```
hduser@superwizard7-VirtualBox:~$ hdfs dfs -ls /lab1/
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 reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/12/14 15:40:48 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 hduser supergroup         89 2020-12-14 15:36 /lab1/input.txt
drwxr-xr-x   - hduser supergroup          0 2020-12-14 15:36 /lab1/output
```

4. Running the JAR file

```
$ hadoop jar /home/superwizard7/Downloads/lab1/wordcount.jar
WordCount /lab1/input.txt /lab1/output/
```

```
hduser@superwizard7-VirtualBox:~$ hadoop jar /home/superwizard7/Downloads/lab1/wordcount.jar WordCount /lab1/input.txt /lab1/output
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 reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/12/14 15:36:49 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
20/12/14 15:36:49 INFO Configuration.deprecation: session.id is deprecated. Instead, use dfs.metrics.session-id
20/12/14 15:36:49 INFO jvm.JvmMetrics: Initializing JVM Metrics with processName=JobTracker, sessionId=
20/12/14 15:36:50 INFO input.FileInputFormat: Total input paths to process : 1
20/12/14 15:36:50 INFO mapreduce.JobSubmitter: number of splits:1
20/12/14 15:36:50 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local685343677_0001
20/12/14 15:36:51 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
20/12/14 15:36:51 INFO mapreduce.Job: Running job: job_local685343677_0001
20/12/14 15:36:51 INFO mapred.LocalJobRunner: OutputCommitter set in config null
20/12/14 15:36:51 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
20/12/14 15:36:51 INFO mapred.LocalJobRunner: Waiting for map tasks
20/12/14 15:36:51 INFO mapred.LocalJobRunner: Starting task: attempt_local685343677_0001_m_000000_0
20/12/14 15:36:51 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
20/12/14 15:36:51 INFO mapred.MapTask: Processing split: hdfs://localhost:54310/lab1/input.txt:0+89
20/12/14 15:36:51 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
20/12/14 15:36:51 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
20/12/14 15:36:51 INFO mapred.MapTask: soft limit at 83886080
20/12/14 15:36:51 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
20/12/14 15:36:51 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
20/12/14 15:36:51 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
20/12/14 15:36:51 INFO mapred.LocalJobRunner:
20/12/14 15:36:51 INFO mapred.MapTask: Starting flush of map output
20/12/14 15:36:51 INFO mapred.MapTask: Spilling map output
20/12/14 15:36:51 INFO mapred.MapTask: bufstart = 0; bufend = 169; bufvoid = 104857600
20/12/14 15:36:51 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214320(104857280); length = 77/6553600
20/12/14 15:36:51 INFO mapred.MapTask: Finished spill 0
20/12/14 15:36:51 INFO mapred.Task: Task:attempt_local685343677_0001_m_000000_0 is done. And is in the process of committing
20/12/14 15:36:51 INFO mapred.LocalJobRunner: map
20/12/14 15:36:51 INFO mapred.Task: Task 'attempt_local685343677_0001_m_000000_0' done.
20/12/14 15:36:51 INFO mapred.LocalJobRunner: Finishing task: attempt_local685343677_0001_m_000000_0
20/12/14 15:36:51 INFO mapred.LocalJobRunner: map task executor complete.
20/12/14 15:36:51 INFO mapred.LocalJobRunner: Waiting for reduce tasks
20/12/14 15:36:51 INFO mapred.LocalJobRunner: Starting task: attempt_local685343677_0001_r_000000_0
20/12/14 15:36:51 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
20/12/14 15:36:51 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@35b5247d
20/12/14 15:36:51 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=375809632, maxSingleShuffleLimit=93952408, mergeThreshold=248034368, ioSortFactor=10, memToMemMergeOutputsThreshold=10
20/12/14 15:36:51 INFO reduce.EventFetcher: attempt_local685343677_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
20/12/14 15:36:51 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local685343677_0001_m_000000_0 decomp: 111 len: 115 to MEMORY
20/12/14 15:36:51 INFO reduce.InMemoryMapOutput: Read 111 bytes from map-output for attempt_local685343677_0001_m_000000_0
20/12/14 15:36:51 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 111, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->111
20/12/14 15:36:51 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
```

5. Output

```
$ hadoop fs –cat /lab1/output/part-r-00000
$ hadoop fs –ls /lab1/output
```

```
hduser@superwizard7-VirtualBox:~$ hdfs dfs -cat /lab1/output/part-r-00000
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.6.0.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 reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/12/14 15:37:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
are     1
brother 1
family  1
hi      1
how     5
is      4
job     1
sister  1
you     1
your    4
hduser@superwizard7-VirtualBox:~$ hdfs dfs -ls /lab1/output
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 reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/12/14 15:37:42 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 hduser supergroup          0 2020-12-14 15:36 /lab1/output/_SUCCESS
-rw-r--r--   1 hduser supergroup         69 2020-12-14 15:36 /lab1/output/part-r-00000
```

6. Stopping Hadoop

```
$ stop-all.sh
```

```
hduser@lenovo-ThinkPad-Edge-E431:~$ stop-all.sh
```

Date: 07-12-2020

6. Hadoop: Average Temperature

Hadoop program to find the Average Temperature

1. Starting Hadoop Cluster

```
$ su hduser
$ cd\
$ start-all.sh
$ jps
```

```
superwizard7@superwizard7-VirtualBox:~$ su hduser
Password:
hduser@superwizard7-VirtualBox:/home/superwizard7$ cd\
>
hduser@superwizard7-VirtualBox:~$ cd /usr/local/hadoop/sbin
hduser@superwizard7-VirtualBox:/usr/local/hadoop/sbin$ ./start-dfs.h
bash: ./start-dfs.h: No such file or directory
hduser@superwizard7-VirtualBox:/usr/local/hadoop/sbin$ ./start-dfs.sh
```

```
hduser@superwizard7-VirtualBox:/usr/local/hadoop/sbin$ jps
8000 ResourceManager
8481 Jps
8356 NodeManager
7813 SecondaryNameNode
7384 NameNode
7582 DataNode
```

2. Copying the binary file to the Hadoop file system as a text file

```
$ hadoop fs -copyFromLocal /home/superwizard7/Downloads/lab2/1901
/lab2/input.txt
$ hadoop –ls /lab2
```

```
hduser@superwizard7-VirtualBox:~$ hdfs dfs -copyFromLocal /home/superwizard7/Downloads/lab2/1901 /lab2/input.txt
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 reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/12/14 15:17:23 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@superwizard7-VirtualBox:~$ hdfs dfs -ls /lab2/
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 reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/12/14 15:17:31 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 hduser supergroup     888190 2020-12-14 15:16 /lab2/input
-rw-r--r--   1 hduser supergroup     888190 2020-12-14 15:17 /lab2/input.txt
```

3. Running the JAR file

```
$ hadoop jar /home/superwizard7/Downloads/lab2/avgtemp.jar AverageDriver
/lab2/input.txt /lab2/output
```

```
hduser@superwizard7-VirtualBox:~$ hadoop jar /home/superwizard7/Downloads/lab2/avgtemp.jar AverageDriver /lab2/input.txt /lab2/output
```

4. Output

```
$ hadoop fs –cat /lab2/output/part-r-00000
$ hadoop fs –ls /lab2/ouput
```

```
hduser@superwizard7-VirtualBox:~$ hdfs dfs -cat /lab2/output/part-r-00000
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.hadoop.security.authentication.util.KerberosUtil (file:/usr/local/hadoop/share/hadoop/common/lib/hadoop-auth-2.6.0.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 reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/12/14 15:20:19 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
1901    46
hduser@superwizard7-VirtualBox:~$ hdfs dfs -ls /lab2/output/
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 reflective access operations
WARNING: All illegal access operations will be denied in a future release
20/12/14 15:20:42 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 hduser supergroup          0 2020-12-14 15:19 /lab2/output/_SUCCESS
-rw-r--r--   1 hduser supergroup          8 2020-12-14 15:19 /lab2/output/part-r-00000
hduser@superwizard7-VirtualBox:~$
```

5. Stopping Hadoop

```
$ stop-all.sh
```

```
hduser@lenovo-ThinkPad-Edge-E431:/home/lenovo$ stop-all.sh
```

Date: 14-12-2020

7. Hive: Employee Table

Write Queries in Hive to do the following

1. **Create an external table named with the following attributes -> Empl_ID ->Emp_Name -> Designation -> Salary**
2. **Load data into table from a given file**
3. **Create a view to Generate a query to retrieve the employee details who earn a salary of more than Rs 30000.**
4. **Alter the table to add a column Dept_Id and Generate a query to retrieve the employee details in order by using Dept_Id**
5. **Generate a query to retrieve the number of employees in each department whose salary is greater than 30000**
6. **Create another table Department with attributes -> Dept_Id ->Dept_name ->Emp_Id**
7. **Display the cumulative details of each employee along with department details**

1. Create an external table named with the following attributes -> Empl_ID ->Emp_Name -> Designation -> Salary

```
>CREATE DATABASE IF NOT EXISTS EMPLOYEES COMMENT ‘EMPLOYEE
Details’ WITH DBPROPERTIES(‘creator’=’Arun’);
>SHOW DATABASES;
>DESCRIBE DATABASE EMPLOYEES;
>USE EMPLOYEES;
> CREATE EXTERNAL TABLE IF NOT EXISTS EMPLOYEES (EMP_ID INT,
EMP_NAME STRING, DESIGNATION STRING, SALARY FLOAT) ROW FORMAT
DELIMITED FIELDS TERMINATED BY '\T' LOCATION '/EMPLOYEE_INFO';
>DESCRIBE FORMATTED EMPLOYEES;
```

2. Load data into table from a given file

```
>INSERT INTO TABLE EMPLOYEES VALUES(1,’Arun’,’Manager’,1000000),
(2,’Ashish’,’Clerk’,50000), (3,’Arvindh’,’Intern’,20000),
(4,’Shruti’,’HR’,35000);
>SELECT * FROM EMPLOYEES;
```

3. Create a view to Generate a query to retrieve the employee details who earn a salary of more than Rs 30000.

```
>CREATE VIEW EMPLOYEE_VIEW AS SELECT * FROM EMPLOYEES WHERE
SALARY>30000;
```

```
>SELECT * FROM EMPLOYEE_VIEW;
```

4. Alter the table to add a column Dept_Id and Generate a query to retrieve the employee details in order by using Dept_Id

```
>ALTER TABLE EMPLOYEES ADD COLUMNS (DEPT_ID INT);
>DESCRIBE FROMATTED EMPLOYEES;
```

5. Generate a query to retrieve the number of employees in each department whose salary is greater than 30000

```
SELECT DEPT_ID, COUNT(DEPT_ID) FROM EMPLOYEES WHERE SALARY >
30000 GROUP BY DEPT_ID;
```

6. Create another table Department with attributes -> Dept_Id ->Dept_name ->Emp_Id

```
CREATE EXTERNAL TABLE IF NOT EXISTS DEPARTMENTS (DEPT_ID INT,
DEPT_NAME STRING, EMP_ID INT) ROW FORMAT DELIMITED FIELDS
TERMINATED BY '\T' LOCATION '/DEPARTMENT';
```

7. Display the cumulative details of each employee along with department details

```
SELECT * FROM EMPLOYEES JOIN DEPARTMENTS ON
EMPLOYEES.DEPT_ID = DEPARTMENTS.DEPT_ID;
```

```
hive> create database if not exists Employees comment 'Employee Details' with dbproperties ('creator'='Arun');
OK
Time taken: 0.179 seconds
hive> show databases;
OK
default
employees
Time taken: 0.022 seconds, Fetched: 2 row(s)
hive> DESCRIBE DATABASE EMPLOYEES
    > DESCRIBE DATABASE EMPLOYEES;
FAILED: ParseException line 2:0 missing EOF at 'DESCRIBE' near 'EMPLOYEES'
hive> DESCRIBE DATABASE EMPLOYEES;
OK
employees       Employee Details        hdfs://localhost:54310/user/hive/warehouse/employees.db hduser  USER
Time taken: 0.056 seconds, Fetched: 1 row(s)
hive> use employees;
OK
Time taken: 0.014 seconds
```

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS EMPLOYEES(EMP_ID INT,EMP_NAME STRING, DESIGNATION STRING, SALARY FLOAT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\T' LOCATION '/EMPLOYEE_INFO';
OK
Time taken: 0.385 seconds
hive> describe formatted employees;
OK
# col_name              data_type               comment

emp_id                  int
emp_name                string
designation             string
salary                  float

# Detailed Table Information
Database:               employees
Owner:                  hduser
CreateTime:             Sat Dec 26 21:00:46 IST 2020
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://localhost:54310/EMPLOYEE_INFO
Table Type:             EXTERNAL_TABLE
Table Parameters:
        EXTERNAL                TRUE
        transient_lastDdlTime   1608996646

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        field.delim             T
        serialization.format    T
Time taken: 0.347 seconds, Fetched: 30 row(s)
hive> insert into table employees values (1,'Arun','Manager',1000000),(2,'Ashish','Clerk',50000),(3,'Arvindh','Intern',20000),(4,'Shruti','HR',35000);
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hduser_20201226210346_a835ef8d-f924-4ae3-a668-53269f851079
Total jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Job running in-process (local Hadoop)
2020-12-26 21:03:51,219 Stage-1 map = 100%,  reduce = 0%
Ended Job = job_local369431830_0001
Stage-4 is selected by condition resolver.
Stage-3 is filtered out by condition resolver.
Stage-5 is filtered out by condition resolver.
```

```
Moving data to directory hdfs://localhost:54310/EMPLOYEE_INFO/.hive-staging_hive_2020-12-26_21-03-46_880_273396099/414858450-1/-ext-10000
Loading data to table employees.employees
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 85 HDFS Write: 253 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 5.2 seconds
hive> select * from employees;
OK
1       Arun    Manager 1000000.0
2       Ashish  Clerk   50000.0
3       Arvindh Intern  20000.0
4       Shruti  HR      35000.0
Time taken: 0.177 seconds, Fetched: 4 row(s)
hive> CREATE VIEW EMPLOYEE_VIEW AS SELECT * FROM EMPLOYEES WHERE SALARY>30000;
OK
Time taken: 0.48 seconds
hive> select * from employee_view;
OK
1       Arun    Manager 1000000.0
2       Ashish  Clerk   50000.0
4       Shruti  HR      35000.0
Time taken: 0.184 seconds, Fetched: 3 row(s)
hive> ALTER TABLE EMPLOYEES ADD COLUMNS (DEPT_ID INT);
OK
Time taken: 0.121 seconds
hive> describe formatted employees;
OK
# col_name              data_type               comment

emp_id                  int
emp_name                string
designation             string
salary                  float
dept_id                 int

# Detailed Table Information
Database:               employees
Owner:                  hduser
CreateTime:             Sat Dec 26 21:00:46 IST 2020
LastAccessTime:         UNKNOWN
Retention:              0
Location:               hdfs://localhost:54310/EMPLOYEE_INFO
Table Type:             EXTERNAL_TABLE
```

```
Table Parameters:
        EXTERNAL                TRUE
        last_modified_by        hduser
        last_modified_time      1608996930
        numFiles                1
        totalSize               93
        transient_lastDdlTime   1608996930

# Storage Information
SerDe Library:          org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe
InputFormat:            org.apache.hadoop.mapred.TextInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        field.delim             T
        serialization.format    T
Time taken: 0.05 seconds, Fetched: 35 row(s)
hive> SELECT DEPT_ID, COUNT(DEPT_ID) FROM EMPLOYEES WHERE SALARY>30000 GROUP BY DEPT_ID;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hduser_20201226210620_28191f63-6e08-45d0-9dda-7ab0c5b48ade
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapreduce.job.reduces=<number>
Job running in-process (local Hadoop)
2020-12-26 21:06:22,239 Stage-1 map = 100%,  reduce = 100%
Ended Job = job_local1509417087_0002
MapReduce Jobs Launched:
Stage-Stage-1:  HDFS Read: 878 HDFS Write: 506 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
NULL    0
Time taken: 1.536 seconds, Fetched: 1 row(s)
hive> CREATE EXTERNAL TABLE IF NOT EXISTS DEPARTMENTS(DEPT_ID INT,DEPT_NAME STRING, EMP_ID INT) ROW FORMAT DELIMITED FIELDS TERMINATED BY '\T' LOCATION '/DEPARTMENT';
OK
Time taken: 0.097 seconds
```

```
hive> SELECT * FROM EMPLOYEES JOIN DEPARTMENTS ON EMPLOYEES.DEPT_ID = DEPARTMENTS.DEPT_ID;
WARNING: Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
Query ID = hduser_20201226212212_6bc64807-5ad0-4e26-9ad7-6f610c2e11b6
Total jobs = 1
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/apache-hive-2.3.7-bin/lib/log4j-slf4j-impl-2.6.2.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
2020-12-26 21:22:17     Starting to launch local task to process map join;      maximum memory = 477626368
2020-12-26 21:22:18     Dump the side-table for tag: 1 with group count: 0 into file: file:/tmp/mydir/503f8a1f-da20-4799-8320-5a20e503d551/hive_2020-12-26_21-22-12_048_5154997920250648288-1/-local-10004/H
ashTable-Stage-3/MapJoin-mapfile11--.hashtable
2020-12-26 21:22:18     Uploaded 1 File to: file:/tmp/mydir/503f8a1f-da20-4799-8320-5a20e503d551/hive_2020-12-26_21-22-12_048_5154997920250648288-1/-local-10004/HashTable-Stage-3/MapJoin-mapfile11--.hasht
able (260 bytes)
2020-12-26 21:22:18     End of local task; Time Taken: 1.186 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Job running in-process (local Hadoop)
2020-12-26 21:22:20,730 Stage-3 map = 100%,  reduce = 0%
Ended Job = job_local374276665_0004
MapReduce Jobs Launched:
Stage-Stage-3:  HDFS Read: 625 HDFS Write: 253 SUCCESS
Total MapReduce CPU Time Spent: 0 msec
OK
Time taken: 8.686 seconds
hive>
```