VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"JnanaSangama", Belgaum -590014, Karnataka.

LAB REPORT
on

BIG DATA ANALYTICS
(20CS6PEBDA)

Submitted by

Jahnavi Satish Shanbhag (1BM19CS065)

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 **Jahnavi Satish Shanbhag(1BM19CS065),** who is 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.

Antara Roy Choudhury
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

1. MongoDB- CRUD Demonstration

```
use
bm65
_db
        switched to db bm65_db


        db.Student.insert({_id:1,name:"Michael",grade:"VII",hobbies:"reading"})

        WriteResult({ "nInserted" : 1 })


        db.Student.update({_id:1},{$set:{hobbies:"cricket"}},{upsert:true})

        WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })


        db.Student.find()

        { "_id" : 1, "name" : "Michael", "grade" : "VII", "hobbies" : "cricket" }


        db.Student.insert({id:1,name:"Latha",grade:"VIII",hobbies:"Singing"})

        WriteResult({ "nInserted" : 1 })


        db.Student.find({name:"Latha"}).pretty()

        {
```

```
        "_id" : ObjectId("6253f120f7936958d67f3c07"),
        "id" : 1,
        "name" : "Latha",
        "grade" : "VIII",
        "hobbies" : "Singing"
}

db.Student.find({},{name:1,grade:1,_id:0})
{ "name" : "Michael", "grade" : "VII" }
{ "name" : "Latha", "grade" : "VIII" }

db.Student.find({grade:{$eq:"VII"}}).pretty()
{ "_id" : 1, "name" : "Michael", "grade" : "VII", "hobbies" : "cricket" }

db.Student.find({name:/^L/}).pretty()
{
        "_id" : ObjectId("6253f120f7936958d67f3c07"),
        "id" : 1,
        "name" : "Latha",
        "grade" : "VIII",
        "hobbies" : "Singing"
}
```

```
db.Student.find({name:/a/}).pretty()
{ "_id" : 1, "name" : "Michael", "grade" : "VII", "hobbies" : "cricket" }
{
        "_id" : ObjectId("6253f120f7936958d67f3c07"),
        "id" : 1,
        "name" : "Latha",
        "grade" : "VIII",
        "hobbies" : "Singing"
}


db.Student.count()
2


db.Student.find().sort({name:1}).pretty()
{
        "_id" : ObjectId("6253f120f7936958d67f3c07"),
        "id" : 1,
        "name" : "Latha",
        "grade" : "VIII",
        "hobbies" : "Singing"
}
```

```
{ "_id" : 1, "name" : "Michael", "grade" : "VII", "hobbies" : "cricket" }
```

```
db.Student.save({name:"Ratan",grade:"VII",_id:1})
```

```
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
```

```
db.Student.find()
```

```
{ "_id" : 1, "name" : "Ratan", "grade" : "VII" }
```

```
{ "_id" : ObjectId("6253f120f7936958d67f3c07"), "id" : 1, "name" : "Latha",
"grade" : "VIII", "hobbies" : "Singing" }
```

```
db.Student.update({_id:1},{$set:{location:"network"}})
```

```
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
```

```
db.Student.update({_id:1},{$unset:{location:"network"}})
```

```
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
```

```
db.Student.find({name:/n$/}).pretty()
```

```
{ "_id" : 1, "name" : "Ratan", "grade" : "VII" }
```

```
db.Student.find({grade:"VII"}).limit(3).pretty()
```

```
{ "_id" : 1, "name" : "Ratan", "grade" : "VII" }
```

```
db.Student.count({grade:"VIII"})
1

db.Student.find().sort({name:1}).pretty()
{
        "_id" : ObjectId("6253f120f7936958d67f3c07"),
        "id" : 1,
        "name" : "Latha",
        "grade" : "VIII",
        "hobbies" : "Singing"
}
{ "_id" : 1, "name" : "Ratan", "grade" : "VII" }

db.Student.find().sort({name:-1}).pretty()
{ "_id" : 1, "name" : "Ratan", "grade" : "VII" }
{
        "_id" : ObjectId("6253f120f7936958d67f3c07"),
        "id" : 1,
        "name" : "Latha",
        "grade" : "VIII",
        "hobbies" : "Singing"
```

```
}



db.Student.find().skip(1).pretty()

{

       "_id" : ObjectId("6253f120f7936958d67f3c07"),

       "id" : 1,

       "name" : "Latha",

       "grade" : "VIII",

       "hobbies" : "Singing"

}



db.createCollection("food")

{ "ok" : 1 }



db.food.insert({_id:1,fruits:['grapes','mango']})

WriteResult({ "nInserted" : 1 })



db.food.insert({_id:2,fruits:['grapes','mango','cherry']})

WriteResult({ "nInserted" : 1 })



db.food.insert({_id:3,fruits:['banana','cherry']})
```

```
WriteResult({ "nInserted" : 1 })
```

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

```
db.food.find({'fruits':{$size:2}})
{ "_id" : 1, "fruits" : [ "grapes", "mango" ] }
{ "_id" : 3, "fruits" : [ "banana", "cherry" ] }
```

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

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

```
db.food.update({_id:3},{$set:{'fruits.1':'apple'}})
WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })
```

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

```
{ "_id" : 2, "fruits" : [ "grapes", "mango", "cherry" ] }

{ "_id" : 3, "fruits" : [ "banana", "apple" ] }



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

WriteResult({ "nMatched" : 1, "nUpserted" : 0, "nModified" : 1 })



db.createCollection("Customers")

{ "ok" : 1 }



db.Customers.insert({custId:1,acctBal:1000,acctType:"current"})

WriteResult({ "nInserted" : 1 })



db.Customers.insert({custId:2,acctBal:2000,acctType:"current"})

WriteResult({ "nInserted" : 1 })



db.Customers.insert({custId:3,acctBal:3000,acctType:"savings"})

WriteResult({ "nInserted" : 1 })



db.Customers.aggregate({$group:{_id:"$custId",toAcctBal:{$sum:"$acctBal"}}}
)

{ "_id" : 3, "toAcctBal" : 3000 }
```

```
{ "_id" : 1, "toAcctBal" : 1000 }

{ "_id" : 2, "toAcctBal" : 2000 }
```

```
db.Customers.aggregate({$match:{acctType:"current"}},{$group:{_id:"$custId"
,toAcctBal:{$sum:"$acctBal"}}})

{ "_id" : 2, "toAcctBal" : 2000 }

{ "_id" : 1, "toAcctBal" : 1000 }
```

```
db.Customers.aggregate({$match:{acctType:"current"}},{$group:{_id:"$custId"
,toAcctBal:{$sum:"$acctBal"}}},{$match:{toAcctBal:{$gt:500}}})

{ "_id" : 2, "toAcctBal" : 2000 }

{ "_id" : 1, "toAcctBal" : 1000 }
```

```
db.Student.drop()

false
```

2. Perform the following DB operations using Cassandra.

1.Create a keyspace by name Employee

2. Create a column family by name

Employee-Info with attributes

Emp_Id Primary Key, Emp_Name,

Designation, Date_of_Joining, Salary,

Dept_Name

3. Insert the values into the table in batch

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

5. Sort the details of Employee records based on salary

6. Alter the schema of the table Employee_Info to add a column Projects which

stores a set of Projects done by the corresponding Employee.

7. Update the altered table to add project names.

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

```
cqlsh> create keyspace
mployee_space WITH
REPLICATION = {'class' :
'SimpleStrategy','replicati
on_factor':2};
```

```
CREATE TABLE employee_space.employee_info (emp_id int
PRIMARY KEY,emp_name text,designation
```

```
text,date_of_joining timestamp,salary float,dept_name
text);



cqlsh> begin batch INSERT INTO
employee_space.employee_info(emp_id,emp_name,designat
ion,date_of_joining,salary,dept_name)
VALUES(1,'Damodar','Manager','2022-01-24',100000,'Mar
keting');

  ... apply batch;

cqlsh> begin batch INSERT INTO
employee_space.employee_info(emp_id,emp_name,designat
ion,date_of_joining,salary,dept_name)
VALUES(2,'Mahalaxmi','Accountant','2021-01-24',200000
,'Accounts');

  ... INSERT INTO
employee_space.employee_info(emp_id,emp_name,designat
ion,date_of_joining,salary,dept_name)
VALUES(3,'Mahesh','Manager','2021-03-24',500000,'Mark
eting');

  ... INSERT INTO
employee_space.employee_info(emp_id,emp_name,designat
ion,date_of_joining,salary,dept_name)
VALUES(4,'Nidhi','Administrator','2021-05-24',500000,
'Administration');

  ... INSERT INTO
employee_space.employee_info(emp_id,emp_name,designat
ion,date_of_joining,salary,dept_name)
VALUES(5,'Rahul','Administrator','2009-05-24',2000000
,'Administration');

  ... apply batch;



cqlsh> use employee_space;
```

```
cqlsh:employee_space> select * from employee_info;

 emp_id | date_of_joining                 | dept_name
| designation    | emp_name  | salary
--------+---------------------------------+----------
------+---------------+-----------+--------
      5 | 2009-05-23 18:30:00.000000+0000 |
Administration | Administrator |     Rahul |  2e+06
      1 | 2022-01-23 18:30:00.000000+0000 |
Marketing |       Manager |   Damodar |  1e+05
      2 | 2021-01-23 18:30:00.000000+0000 |
Accounts |    Accountant | Mahalaxmi |  2e+05
      4 | 2021-05-23 18:30:00.000000+0000 |
Administration | Administrator |     Nidhi |  5e+05
      3 | 2021-03-23 18:30:00.000000+0000 |
Marketing |       Manager |    Mahesh |  5e+05

(5 rows)

cqlsh:employee_space> update employee_info set
emp_name='Radha' where emp_id=1;

cqlsh:employee_space> update employee_info set
dept_name='Development' where emp_id=1;

cqlsh:employee_space> select * from employee_info;
```

```
emp_id | date_of_joining                  | dept_name
| designation    | emp_name  | salary

--------+----------------------------------+----------
------+---------------+-----------+--------

     5 | 2009-05-23 18:30:00.000000+0000 |
Administration | Administrator |     Rahul |  2e+06

     1 | 2022-01-23 18:30:00.000000+0000 |
Development |       Manager |     Radha |  1e+05

     2 | 2021-01-23 18:30:00.000000+0000 |
Accounts |    Accountant | Mahalaxmi |  2e+05

     4 | 2021-05-23 18:30:00.000000+0000 |
Administration | Administrator |     Nidhi |  5e+05

     3 | 2021-03-23 18:30:00.000000+0000 |
Marketing |       Manager |    Mahesh |  5e+05



(5 rows)
```

```
cqlsh:employee_space> alter table employee_info add
projects set<text>;
```

```
cqlsh:employee_space> update employee_info set
projects=projects+{'Web development','machine
learning'} where emp_id=2;
```

```
cqlsh:employee_space> select * from employee_info;
```

```
emp_id | date_of_joining                 | dept_name
| designation   | emp_name  | projects
| salary

--------+---------------------------------+----------
------+---------------+-----------+-----------------
----------------------+--------

     5 | 2009-05-23 18:30:00.000000+0000 |
Administration | Administrator |     Rahul |
null |  2e+06

     1 | 2022-01-23 18:30:00.000000+0000 |
Development |       Manager |     Radha |
null |  1e+05

     2 | 2021-01-23 18:30:00.000000+0000 |
Accounts |    Accountant | Mahalaxmi | {'Web
development', 'machine learning'} |  2e+05

     4 | 2021-05-23 18:30:00.000000+0000 |
Administration | Administrator |     Nidhi |
null |  5e+05

     3 | 2021-03-23 18:30:00.000000+0000 |
Marketing |       Manager |    Mahesh |
null |  5e+05



(5 rows)

cqlsh:employee_space> update employee_info set
projects=projects+{'Web development','machine
learning','cybersecurity'} where emp_id=5;



cqlsh:employee_space> select * from employee_info;



emp_id | date_of_joining                 | dept_name
| designation   | emp_name  | projects
| salary
```

```
--------+----------------------------------+----------
------+---------------+-----------+------------------
---------------------------------------+--------

     5 | 2009-05-23 18:30:00.000000+0000 |
Administration | Administrator |     Rahul | {'Web
development', 'cybersecurity', 'machine learning'} |
2e+06

     1 | 2022-01-23 18:30:00.000000+0000 |
Development |       Manager |     Radha |
null |  1e+05

     2 | 2021-01-23 18:30:00.000000+0000 |
Accounts |    Accountant | Mahalaxmi |
{'Web development', 'machine learning'} |  2e+05

     4 | 2021-05-23 18:30:00.000000+0000 |
Administration | Administrator |     Nidhi |
null |  5e+05

     3 | 2021-03-23 18:30:00.000000+0000 |
Marketing |       Manager |    Mahesh |
null |  5e+05



(5 rows)



cqlsh:employee_space> INSERT INTO
employee_space.employee_info(emp_id,emp_name,designat
ion,date_of_joining,salary,dept_name)
VALUES(6,'Harshitha','Manager','2022-01-24',100000,'M
arketing') using ttl 15;



cqlsh:employee_space> select * from employee_info;
```

```
emp_id | date_of_joining                 | dept_name
| designation   | emp_name  | projects
| salary

--------+---------------------------------+----------
------+--------------+-----------+-----------------
---------------------------------------+--------

     5 | 2009-05-23 18:30:00.000000+0000 |
Administration | Administrator |     Rahul | {'Web
development', 'cybersecurity', 'machine learning'} |
2e+06

     1 | 2022-01-23 18:30:00.000000+0000 |
Development |       Manager |     Radha |
null |  1e+05

     2 | 2021-01-23 18:30:00.000000+0000 |
Accounts |    Accountant | Mahalaxmi |
{'Web development', 'machine learning'} |  2e+05

     4 | 2021-05-23 18:30:00.000000+0000 |
Administration | Administrator |     Nidhi |
null |  5e+05

     6 | 2022-01-23 18:30:00.000000+0000 |
Marketing |       Manager | Harshitha |
null |  1e+05

     3 | 2021-03-23 18:30:00.000000+0000 |
Marketing |       Manager |    Mahesh |
null |  5e+05



(6 rows)



cqlsh:employee_space> select * from employee_info;



emp_id | date_of_joining                 | dept_name
| designation   | emp_name  | projects
| salary
```

```
--------+---------------------------------+----------
------+---------------+-----------+------------------
---------------------------------------+--------

     5 | 2009-05-23 18:30:00.000000+0000 |
Administration | Administrator |     Rahul | {'Web
development', 'cybersecurity', 'machine learning'} |
2e+06

     1 | 2022-01-23 18:30:00.000000+0000 |
Development |       Manager |     Radha |
null |  1e+05

     2 | 2021-01-23 18:30:00.000000+0000 |
Accounts |    Accountant | Mahalaxmi |
{'Web development', 'machine learning'} |  2e+05

     4 | 2021-05-23 18:30:00.000000+0000 |
Administration | Administrator |     Nidhi |
null |  5e+05

     3 | 2021-03-23 18:30:00.000000+0000 |
Marketing |       Manager |    Mahesh |
null |  5e+05



(5 rows)
```

3. Perform the following DB operations using Cassandra.

1.Create a keyspace by name Library

2. Create a column family by name Library-Info with attributes

Stud_Id Primary Key,

Counter_value of type Counter,

Stud_Name, Book-Name, Book-Id,

Date_of_issue

3. Insert the values into the table in batch

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

5. Write a query to show that a student with id 112 has taken a book “BDA” 2 times.

6. Export the created column to a csv file

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

```
cqlsh> create keyspace
library_space WITH
REPLICATION={'class':'Sim
pleStrategy','replication
_factor':2};
```

```
cqlsh> use library_space;
```

```
cqlsh:library_space> create table library_info(stud_id
int,counter_value counter,stud_name text,book_name
text,book_id int,date_of_issue timestamp,PRIMARY
KEY(stud_id,stud_name,book_name,book_id,date_of_issue))
;
```

```
cqlsh:library_space> update library_info set
counter_value=counter_value+1 where stud_id=1 and
stud_name='abc' and book_name='book1' and book_id=11
and date_of_issue='2022-01-30';
```

```
cqlsh:library_space> update library_info set
counter_value=counter_value+1 where stud_id=2 and
stud_name='def' and book_name='book2' and book_id=12
and date_of_issue='2022-03-30';
```

```
cqlsh:library_space> update library_info set
counter_value=counter_value+1 where stud_id=3 and
stud_name='ghi' and book_name='book3' and book_id=13
and date_of_issue='2022-05-30';
```

```
cqlsh:library_space> update library_info set
counter_value=counter_value+1 where stud_id=4 and
stud_name='jkl' and book_name='book4' and book_id=14
and date_of_issue='2022-07-30';
```

```
cqlsh:library_space> update library_info set
counter_value=counter_value+1 where stud_id=5 and
stud_name='mno' and book_name='book5' and book_id=15
and date_of_issue='2022-09-30';
```

```
cqlsh:library_space> select * from library_info;

 stud_id | stud_name | book_name | book_id |
date_of_issue                   | counter_value
---------+-----------+-----------+---------+-----------
---------------------+---------------

       5 |       mno |     book5 |      15 | 2022-09-29
18:30:00.000000+0000 |             1

       1 |       abc |     book1 |      11 | 2022-01-29
18:30:00.000000+0000 |             1

       2 |       def |     book2 |      12 | 2022-03-29
18:30:00.000000+0000 |             1

       4 |       jkl |     book4 |      14 | 2022-07-29
18:30:00.000000+0000 |             1

       3 |       ghi |     book3 |      13 | 2022-05-29
18:30:00.000000+0000 |             1


(5 rows)


cqlsh:library_space> update library_info set
counter_value=counter_value+1 where stud_id=5 and
stud_name='mno' and book_name='book5' and book_id=15
and date_of_issue='2022-09-30';


cqlsh:library_space> select * from library_info;
```

```
stud_id | stud_name | book_name | book_id |
date_of_issue                    | counter_value

---------+-----------+-----------+---------+-----------
----------------------+---------------

      5 |       mno |     book5 |      15 | 2022-09-29
18:30:00.000000+0000 |             2

      1 |       abc |     book1 |      11 | 2022-01-29
18:30:00.000000+0000 |             1

      2 |       def |     book2 |      12 | 2022-03-29
18:30:00.000000+0000 |             1

      4 |       jkl |     book4 |      14 | 2022-07-29
18:30:00.000000+0000 |             1

      3 |       ghi |     book3 |      13 | 2022-05-29
18:30:00.000000+0000 |             1



(5 rows)



cqlsh:library_space> copy
library_info(stud_id,stud_name,book_name,book_id,date_o
f_issue,counter_value) to
'/home/bmscecse/Desktop/bda.csv';



Using 11 child processes



Starting copy of library_space.library_info with
columns [stud_id, stud_name, book_name, book_id,
date_of_issue, counter_value].

Processed: 5 rows; Rate:      45 rows/s; Avg. rate:
45 rows/s

5 rows exported to 1 files in 0.121 seconds.
```

```
cqlsh:library_space> create table
library_info_copy(stud_id int,counter_value
counter,stud_name text,book_name text,book_id
int,date_of_issue timestamp,PRIMARY
KEY(stud_id,stud_name,book_name,book_id,date_of_issue))
;
```

```
cqlsh:library_space> copy
library_info_copy(stud_id,stud_name,book_name,book_id,d
ate_of_issue,counter_value) from
'/home/bmscecse/Desktop/new.csv';
```

```
Using 11 child processes
```

```
Starting copy of library_space.library_info_copy with
columns [stud_id, stud_name, book_name, book_id,
date_of_issue, counter_value].

Processed: 5 rows; Rate:       8 rows/s; Avg. rate:
12 rows/s

5 rows imported from 1 files in 0.406 seconds (0
skipped).
```

```
cqlsh:library_space> select * from library_info where
counter_value=2 allow filtering;
```

```
stud_id | stud_name | book_name | book_id |
date_of_issue                   | counter_value
```

```
---------+-----------+-----------+---------+-----------
----------------------+---------------

      5 |       mno |     book5 |      15 | 2022-09-29
18:30:00.000000+0000 |             2
```