

VISVESVARAYA TECHNOLOGICAL UNIVERSITY

“JnanaSangama”, Belgaum -590014, Karnataka.



LAB REPORT on

Big Data Analytics(23CS6PCBDA)

Submitted by

Mohammed Shuraim (1BM22CS158)

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
Feb-2025 to June-2025

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(23CS6PCBDA)**” carried out by **Mohammed Shuraim (1BM22CS158)**, 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 2024. The Lab report has been approved as it satisfies the academic requirements in respect of a **Course Title - (Course code)** work prescribed for the said degree.

Amruta B
Assistant Professor
Department of CSE
BMSCE, Bengaluru

Dr. Kavitha Sooda
Professor and Head
Department of CSE
BMSCE, Bengaluru

Index Sheet

Sl. No.	Experiment Title	Page No.
1	MongoDB	4
2	MongoDB	6
3	Neo4j	14
4	Cassandra	24
5	Cassandra	32
6	Hadoop	40
7	Hadoop	44
8	Hadoop	50
8	Scala and Spark	55

Course Outcome

CO1	Apply the concept of NoSQL, Hadoop or Spark for a given task
CO2	Analyse big data analytics mechanisms that can be applied to obtain solution for a given problem.
CO3	Design and implement solutions using data analytics mechanisms for a given problem.

Lab 1: MongoDB- CRUD Demonstration

Question: Perform basic CRUD (Create, Read, Update, Delete) operations in MongoDB.

Code with Output:

LAB 1 - 18M22CS158 - BDA MongoDB commands

```

def startMethod() {
    // 1. Create a new instance of the class
    // 2. Call the start() method
    // 3. Return the result
    // 4. End of the function
}

// 1. Create a new instance of the class
// 2. Call the start() method
// 3. Return the result
// 4. End of the function

```

```

Atlas atlas-shard-0 [primary] mydb> db.Student.insert([{"_id":"1", "StudentName":"MichelleSachintha", "Grade": "V", "Hobbies":["InterestSurfing"]}],
DeprecationWarning: Collection.insert() is deprecated. Use insertOne, insertMany, or bulkWrite.
{ acknowledged: true, insertedIds: { '0': '1' } }
Atlas atlas-shard-0 [primary] mydb> db.Student.insertMany([{"_id":"2", "StudentName":"JohnCanny", "grade": "V", "Hobbies":["Blogging"]}, {"_id":"3", "StudentName":"JaneSmith", "grade":"V", "Hobbies":["graphicDesign"]}],
{ acknowledged: true, insertedIds: { '0': '2', '1': '3' } }
Atlas atlas-shard-0 [primary] mydb> db.Student.find()
{
  "_id": "1",
  "StudentName": "MichelleSachintha",
  "Grade": "V",
  "Hobbies": "InterestSurfing",
},
{ "_id": "2", "StudentName": "JohnCanny", "grade": "V", "Hobbies": "Blogging" },
{
  "_id": "3",
  "StudentName": "JaneSmith",
  "grade": "V",
  "Hobbies": "graphicDesign"
}
Atlas atlas-shard-0 [primary] mydb> db.Student.find({}, {StudentName:1, Grade:1, _id:0}),
{
  StudentName: "MichelleSachintha", Grade: "V" },
{
  StudentName: "JohnCanny",
},
{
  StudentName: "JaneSmith"
}
Atlas atlas-shard-0 [primary] mydb> db.Student.find({}, {StudentName:1, Grade:1, _id:0});
{
  StudentName: "MichelleSachintha", Grade: "V" },
{
  StudentName: "JohnCanny", Grade: "V" },
{
  StudentName: "JaneSmith", Grade: "V" }
Atlas atlas-shard-0 [primary] mydb> db.Student.find({}, {StudentName:1, Grade:1, _id:0});
{
  StudentName: "MichelleSachintha", Grade: "V" },
{
  StudentName: "JohnCanny",
},
{
  StudentName: "JaneSmith"
}
Atlas atlas-shard-0 [primary] mydb>

```

```
PS C:\Users\student> mongoexport mongodb+srv://shuraimcs22:Mshuraim1010@cluster0.p
2025-03-11T15:45:09.598+0530    connected to: mongodb+srv://shuraimcs22:Mshuraim10
2025-03-11T15:45:10.128+0530    exported 6 records
PS C:\Users\student> mongoimport mongodb+srv://shuraimcs22:Mshuraim1010@cluster0.p
2025-03-11T15:47:34.696+0530    connected to: mongodb+srv://shuraimcs22:Mshuraim10
2025-03-11T15:47:34.838+0530    6 document(s) imported successfully. 0 document(s)
PS C:\Users\student>
```

4/3/25

store
67

LAB-1

Export / Import the created table into
local file system

Database ~~1478~~
dev-class

Export:

```
PS C:\Users\student> mongoexport mongodb+srv:  
//shuramcs22:IMShuram101@cluster0.paul  
mongodb.net/dbms_demo --collection=  
Student --out C:\user\student\devlop\bu  
son.json
```

```
2025-03-04T15:20:09.598+0530  
connected to: mongodb+srv://S**Redacted]  
/dbms_demo
```

Exported 6 records

```
PS C:\user\student> mongorestore mongo+srv:  
//shuramcs22:IMShuram101@cluster0.paul  
mongodb.net/dbms_demo --collection=Student  
--type --file C:\user\student\devlop\bu  
son.json
```

~~6~~
~~2153~~

Lab 2: MongoDB- CRUD Demonstration

Question: Perform basic CRUD (Create, Read, Update, Delete) operations in MongoDB.

Code with Output:

LAB 2 - 1BM22CS158 - BDA MongoDB commands

```
Terminal
MongoDB Shell
mongo --host 127.0.0.1:27017 --directConnection=true --serverSelectionTimeoutMS=2000

Successfully installed - HP-Elite-Tower-800-G9-Desktop-PC: $ brew services start mongodb-community
Command 'brew' not found, did you mean:
  command 'qbrew' from deb qbrew (0.4.1-0build1)
  command 'brec' from deb bplay (0.991-10build1)
Try: sudo apt install <deb name>
Successfully installed - HP-Elite-Tower-800-G9-Desktop-PC: $ sh
$ mongosh

Current Mongosh Log ID: 67c7f94fd36ae4456efe6938
Connecting to:
  mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000&appName=mongosh-2.3.2
Using MongoDB:
  7.0.14
Using Mongosh:
  2.3.2
Mongosh 2.4.2 is available for download: https://www.mongodb.com/try/download/shell

For mongosh info see: https://www.mongodb.com/docs/mongosh-shell/

*****
The server generated these startup warnings when booting:
2021-03-11T14:11:24.613+05:30: Using the XFS filesystem is strongly recommended with the WiredTiger storage engine. See http://dochub.mongodb.org/core/production-file-system
2021-03-11T14:11:28.388+05:30: Access control is not enabled for the database. Read and write access to data and configuration is unrestricted
*****

test>
```

```
test> use myDB;
switched to db myDB
myDB> db;
myDB
myDB> show dbs;
admin      40.00 KiB
config     72.00 KiB
local      128.00 KiB
mydb       40.00 KiB
myDB> db.createCollection("Student");
MongoServerError[DatabaseDifferCase]: db already exists with different case already have: [mydb] trying to create [myDB]
myDB> use shdb
switched to db shdb
shdb>
```

```

myDB> use shdb
switched to db shdb
shdb> db.createCollection("Student");
{ ok: 1 }
shdb> db.Student.drop();
true
shdb> db.createCollection("Student");
{ ok: 1 }
shdb> db.Student.insert({_id:1, StudName:"MichelleJacintha", Grade:"VII", Hobbies:"InternetSurfing"});
DeprecationWarning: Collection.insert() is deprecated. Use insertOne, insertMany, or bulkWrite.
{ acknowledged: true, insertedIds: { '0': 1 } }
shdb> db.Student.update(
...   { _id:3, StudName:"AryanDavid", Grade:"VII"},
...   { $set:{Hobbies:"Skating"}},
...   { upsert:true }
... );
DeprecationWarning: Collection.update() is deprecated. Use updateOne, updateMany, or bulkWrite.
{
  acknowledged: true,
  insertedId: 3,
  matchedCount: 0,
  modifiedCount: 0,
  upsertedCount: 1
}
shdb> db.Student.find({StudName:"Aryan David"});

shdb> db.Student.find({StudName:"Aryan David"});

shdb> db.Student.find({}, {StudName:1, Grade:1, _id:0});
[
  { StudName: 'MichelleJacintha', Grade: 'VII' },
  { Grade: 'VII', StudName: 'AryanDavid' }
]
shdb> db.Student.find({StudName:"Aryan David"});

shdb> db.Student.find();
[
  {
    _id: 1,
    StudName: 'MichelleJacintha',
    Grade: 'VII',
    Hobbies: 'InternetSurfing'
  },
  { _id: 3, Grade: 'VII', StudName: 'AryanDavid', Hobbies: 'Skating' }
]
shdb> db.Student.find({StudName:"Arya David"});

shdb> db.Student.find({StudName:"AryaDavid"});

shdb> db.Student.find({}, {StudName:1, Grade:1, _id:0});
[
  { StudName: 'MichelleJacintha', Grade: 'VII' },
  { Grade: 'VII', StudName: 'AryanDavid' }
]
shdb> █

```



```
shdb> db.Student.find({Grade:{Seq: 'VII'}}).pretty();
[
  {
    _id: 1,
    StudName: 'MichelleJacintha',
    Grade: 'VII',
    Hobbies: 'InternetSurfing'
  },
  { _id: 3, Grade: 'VII', StudName: 'AryanDavid', Hobbies: 'Skating' }
]
shdb> db.Student.find({Hobbies: { $in: ['Chess','Skating'] }}).pretty();
[
  { _id: 3, Grade: 'VII', StudName: 'AryanDavid', Hobbies: 'Skating' }
]
shdb> db.Student.find({StudName:/M/}).pretty();
[
  {
    _id: 1,
    StudName: 'MichelleJacintha',
    Grade: 'VII',
    Hobbies: 'InternetSurfing'
  }
]
shdb> db.Student.find({StudName:/a/}).pretty();
[
  {
    _id: 1,
    StudName: 'MichelleJacintha',
    Grade: 'VII',
    Hobbies: 'InternetSurfing'
  }
]
shdb> db.Student.count();
DeprecationWarning: Collection.count() is deprecated. Use countDocuments or estimatedDocumentCount.
2
shdb> db.Student.find().sort({StudName:-1}).pretty();
[
  {
    _id: 1,
    StudName: 'MichelleJacintha',
    Grade: 'VII',
    Hobbies: 'InternetSurfing'
  },
  { _id: 3, Grade: 'VII', StudName: 'AryanDavid', Hobbies: 'Skating' }
]
shdb>
```

[illegible]

```

2025-03-17T13:04:46.831+0300 try 'mongoexport --help' for more information
mongoexport --uri="mongodb://127.0.0.1:27010" --host localhost --db shdb --collection student --type=csv --out /home/bmccccc/Desktop/output.txt --fields StudName,Address
2025-03-17T13:04:46.847+0300 connected to: mongoexport://localhost/
mongoexport --uri="mongodb://127.0.0.1:27010" --host localhost --db shdb --collection student --type=csv --out /home/bmccccc/Desktop/output.txt --fields StudName,Address
2025-03-17T13:04:46.861+0300 connected to: mongoexport://localhost/
2025-03-17T13:04:46.875+0300 exported 3 records
mongoexport --uri="mongodb://127.0.0.1:27010" --host localhost --db shdb --collection student --type=csv --headerline --file /home/bmccccc/Desktop/students.csv
2025-03-17T13:04:46.889+0300 Failed open /home/bmccccc/Desktop/students.csv: no such file or directory
2025-03-17T13:04:46.903+0300 0 document(s) imported successfully, 0 document(s) failed to import.
mongoexport --uri="mongodb://127.0.0.1:27010" --host localhost --db shdb --collection student --type=csv --headerline --file /home/bmccccc/Desktop/students.csv
2025-03-17T13:04:46.917+0300 Failed open /home/bmccccc/Desktop/students.csv: no such file or directory
mongoexport --uri="mongodb://127.0.0.1:27010" --host localhost --db shdb --collection student --type=csv --headerline --file /home/bmccccc/Desktop/students.csv
2025-03-17T13:04:46.931+0300 Failed open /home/bmccccc/Desktop/students.csv: no such file or directory
mongoexport --uri="mongodb://127.0.0.1:27010" --host localhost --db shdb --collection student --type=csv --headerline --file /home/bmccccc/Desktop/students.csv
2025-03-17T13:04:46.945+0300 connected to: mongoexport://localhost/
2025-03-17T13:04:46.959+0300 continuing through error: E11000 duplicate key error collection: shdb.student index: _id, dup key: { "_id": 1 }
2025-03-17T13:04:46.973+0300 continuing through error: E11000 duplicate key error collection: shdb.student index: _id, dup key: { "_id": 2 }
2025-03-17T13:04:46.987+0300 continuing through error: E11000 duplicate key error collection: shdb.student index: _id, dup key: { "_id": 3 }
2025-03-17T13:04:46.999+0300 3 document(s) imported successfully, 3 document(s) failed to import.
mongoexport --uri="mongodb://127.0.0.1:27010" --host localhost --db shdb --collection student --type=csv --headerline --file /home/bmccccc/Desktop/students.csv

```

```

TypeError: db.Students.save is not a function
test> db.Students.update({_id:4}, {$set:{Location:"Network"}});
DeprecationWarning: Collection.update() is deprecated. Use updateOne, updateMany, or bulkWrite.
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 0,
  modifiedCount: 0,
  upsertedCount: 0
}
test> db.Students.update({_id:4}, {$unset:{Location:"Network"}});
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 0,
  modifiedCount: 0,
  upsertedCount: 0
}
test> db.Students.update({_id:4}, {$unset:{Location:"Network"}});
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 0,
  modifiedCount: 0,
  upsertedCount: 0
}
test> db.Students.update({_id:3}, {$set:{Location:null}});
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 0,
  modifiedCount: 0,
  upsertedCount: 0
}
test> db.Students.find({Grade:"VII"}).limit(3).pretty();
test> db.Students.find().sort({StudName:1}).pretty();

test> db.food.insert({_id:1, fruits:['grapes','mango','apple']});
DeprecationWarning: Collection.insert() is deprecated. Use insertOne, insertMany, or bulkWrite.
{ acknowledged: true, insertedIds: { '0': 1 } }
test> db.food.insert({_id:2, fruits:['grapes','mango','cherry']});
{ acknowledged: true, insertedIds: { '0': 2 } }
test> db.food.insert({_id:3, fruits:['banana','mango']});
{ acknowledged: true, insertedIds: { '0': 3 } }
test> db.food.find({fruits: ['grapes','mango','apple']}).pretty();
[ { _id: 1, fruits: [ 'grapes', 'mango', 'apple' ] },
  { _id: 2, fruits: [ 'grapes', 'mango', 'cherry' ] },
  { _id: 3, fruits: [ 'banana', 'mango' ] } ]
test> db.food.find({'fruits.1':'grapes'});
test> db.food.find({'fruits.1':'grapes'});

```

```

test> db.food.find({"fruits": {$size:2}});
[ { _id: 3, fruits: [ 'banana', 'mango' ] } ]
test> db.food.find({_id:1}, {"fruits":{$slice:2}});
[ { _id: 1, fruits: [ 'grapes', 'mango' ] } ]
test> db.food.update({_id:3}, {$set: {'fruits.1':'apple'}});
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
test> db.food.update({_id:2}, {$push: {price:{grapes:80,mango:200,cherry:100}}});
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
test> db.food.update({_id:3}, {$set: {'fruits.1':'apple'}});
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 0,
  upsertedCount: 0
}
test> db.Customers.aggregate([{$group: { _id: "$custID", TotAccBal: {$sum:"$AcctBal"} } }]);

test> db.Customers.aggregate([
...   {$match:{AcctType:"S"} },
...   {$group: { _id: "$custID", TotAccBal: {$sum:"$AcctBal"} } }
... ]);

test> db.Customers.aggregate([
...   {$match:{AcctType:"S"} },
...   {$group: { _id: "$custID", TotAccBal: {$sum:"$AcctBal"} } },
...   {$match:{TotAccBal:{$gt:1200}}}
... ]);

test> db.Alphabets.insertMany([{_id:1, alphabet:"A"}, {_id:2, alphabet:"B"}, {_id:3, alphabet:"C"}]);
{ acknowledged: true, insertedIds: { '0': 1, '1': 2, '2': 3 } }
test> var myCursor = db.Alphabets.find();

```

```

    matchedCount: 1,
    modifiedCount: 1,
    upsertedCount: 0
  }
}
test> db.food.update({_id:3}, {$set: {'fruits.1':'apple'}});
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 0,
  upsertedCount: 0
}
test> db.Customers.aggregate([{$group: { _id: "$custID", TotAccBal: {$sum:"$AcctBal"} } }]);

test> db.Customers.aggregate([
...   {$match:{AcctType:"S"} },
...   {$group: { _id: "$custID", TotAccBal: {$sum:"$AcctBal"} } }
... ]]);

test> db.Customers.aggregate([
...   {$match:{AcctType:"S"} },
...   {$group: { _id: "$custID", TotAccBal: {$sum:"$AcctBal"} } },
...   {$match:{TotAccBal:{$gt:1200}}}
... ]]);

test> db.Alphabets.insertMany([{$_id:1, alphabet:"A"}, {$_id:2, alphabet:"B"}, {$_id:3, alphabet:"C"}]);
{ acknowledged: true, insertedIds: { '0': 1, '1': 2, '2': 3 } }
test> var myCursor = db.Alphabets.find();

test> while (myCursor.hasNext()) {
...   printjson(myCursor.next());
... }
{
  _id: 1,
  alphabet: 'A'
}
{
  _id: 2,
  alphabet: 'B'
}
{
  _id: 3,
  alphabet: 'C'
}

test> show dbs;
admin    40.00 KiB
config  108.00 KiB
local   128.00 KiB
nydb     40.00 KiB
shdb    112.00 KiB
test     96.00 KiB
test>

```


Lab 2: MongoDB CRUD operations

> use shdb

```
shdb> use shdb
shdb> db.createCollection("Student")
{ ok: 1 }
```

```
shdb> db.Student.insert({ _id: 1, studentname:
  "Michelle Jacinto", grade: "VII", Hobbies: "Internet
  Surfing" });
{ acknowledged: true, insertedIds: { "0": 1 } }
```

```
shdb> db.Student.update({ _id: 3
  studentname: "Aryen David", grade: "VII" },
  { $set: { Hobbies: "skating" } },
  { upsert: true }
);
```

```
shdb> db.Student.find({ studentname: 1, grade:
  "VII" },
  { _id: 0 });
[
  { studentname: 'Michelle Jacinto', grade: 'VII' },
  { grade: 'VII', studentname: 'Aryen David' }
]
```

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

Composed } ~~source~~ ^{written} in Lab 1

store
67

```
test > db.Students.update({-id: 4},
  { $set: { location: 'Netherlands' } },
  {
    acknowledged: true,
    inserted: null,
    matchedCount: 0,
    modifiedCount: 0,
    upsertedCount: 0
  })
```

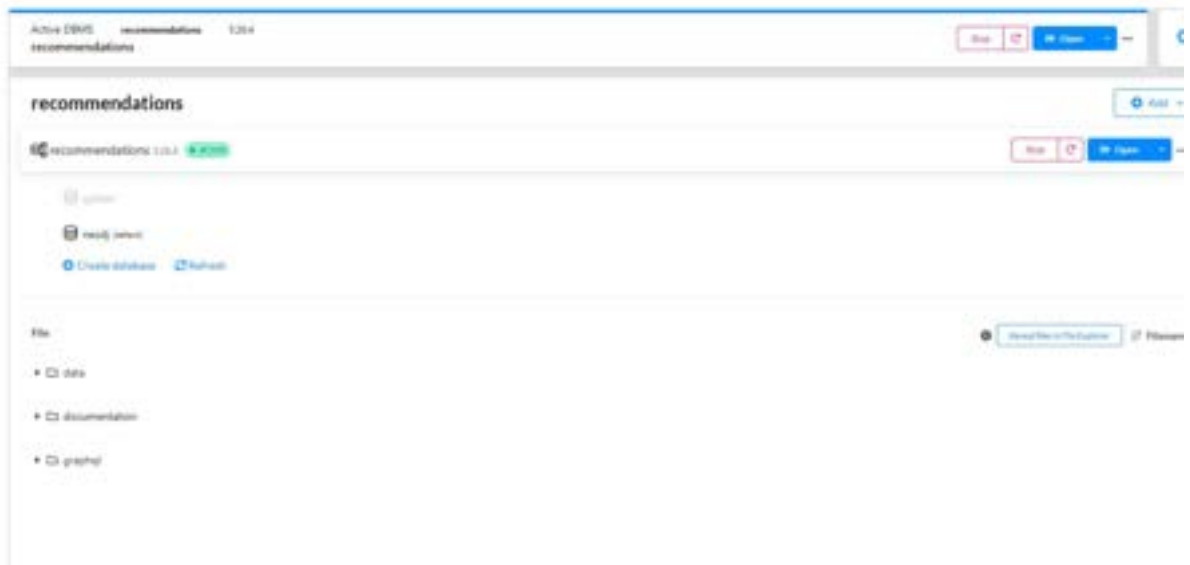
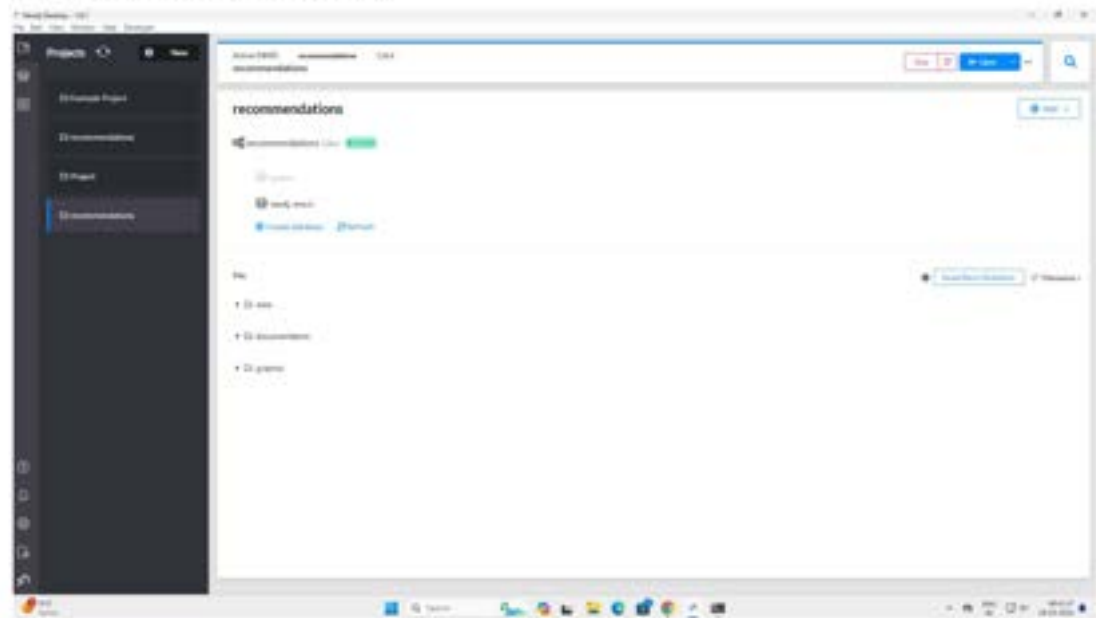
```
test > db.food.insert({-id: 1, fruits: ['grapes',
  ['grapes', 'mango', 'apple']] });
{ acknowledged: true; inserted: { '0': 1 } }
```

```
test > db.customers.aggregate({ $group:
  { _id: $awt10, TotalAccBal: { $sum: '$AccBal' } } })
```

```
test > db.Customers.aggregate({
  { $match: { AccType: 'S' } },
  { $group: { _id: '$awt10', TotalAccBal: { $sum:
    '$AccBal' } } },
  { $match: { TotalAccBal: { $gt: 1200 } } }
})
```

Lab 3: Neo4J

LAB 3 - 18M22CS158 - BDA -Neo 4j

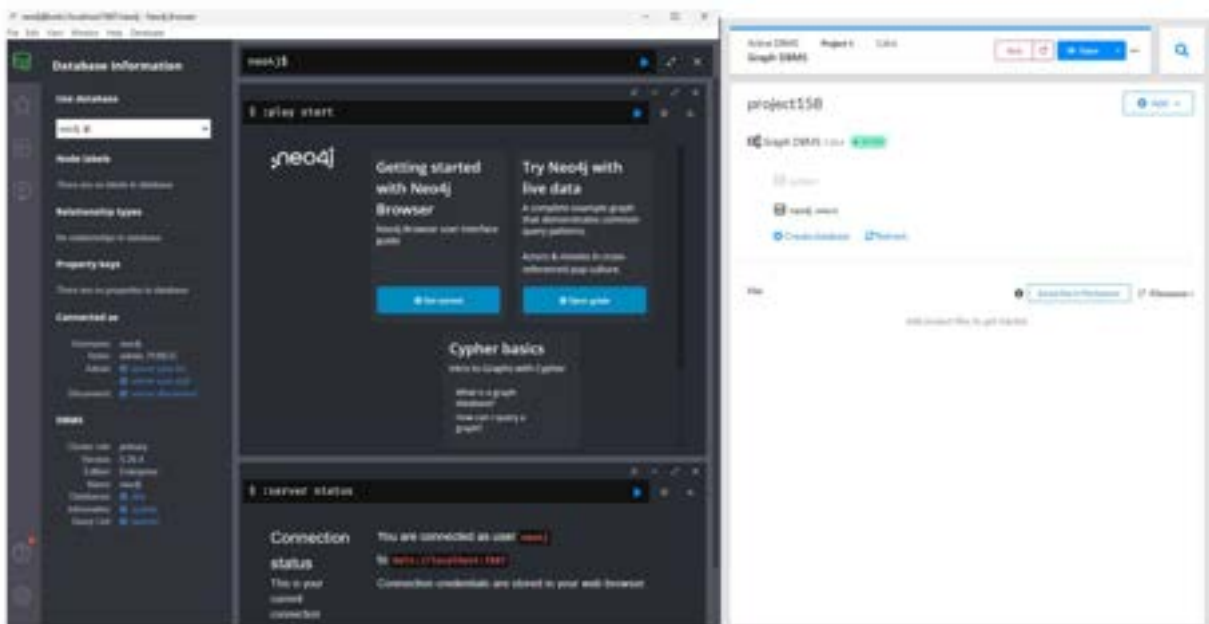
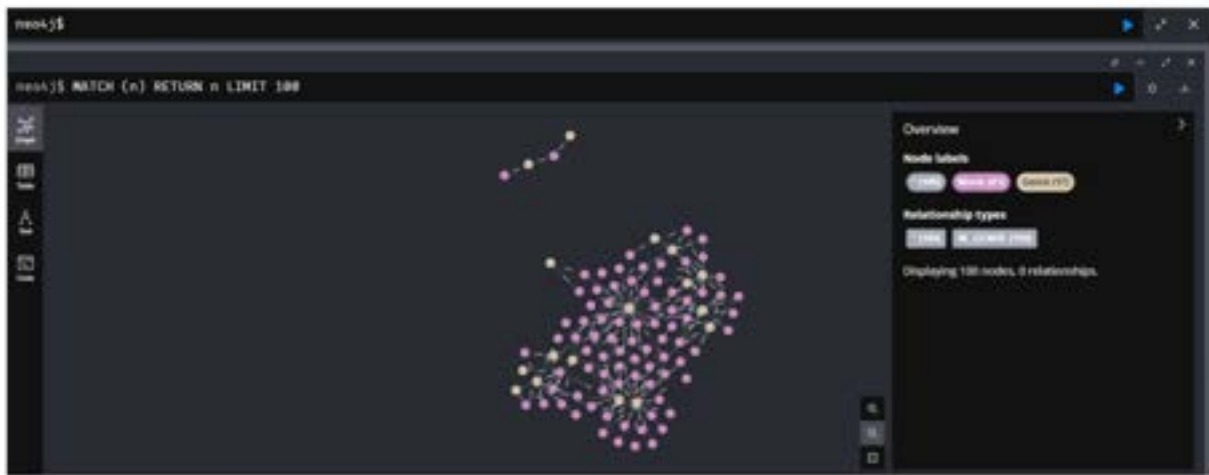


Neo4j Desktop interface showing the Neo4j Browser. The left sidebar displays "Database Information" for the "neo4j" database, including node labels (Genre, Director, Actor, Person, User) and relationship types (DIRECTED, ACTED_IN, RATED). The main area shows a graph visualization of the schema, with nodes representing different roles and relationships between them. The graph structure is as follows:

```
graph TD; Genre((Genre)) -->|IN_GENRE| Movies((Movies)); Director((Director)) -->|DIRECTED| Movies; Actor((Actor)) -->|ACTED_IN| Movies; Person((Person)) -->|DIRECTED| Movies; Person -->|ACTED_IN| Movies; User((User)) -->|RATED| Movies;
```

The right sidebar shows a "Overview" section with the same node labels and relationship types. Below the graph, there are links to "Getting started with Neo4j Browser", "Try Neo4j with live data", and "Cypher basics".







```
neo4j> CREATE (a: STUDENT {name:"Rago",id:11})
```

Added 1 node, created 1 node, and 0 properties, completed after 1 ms.

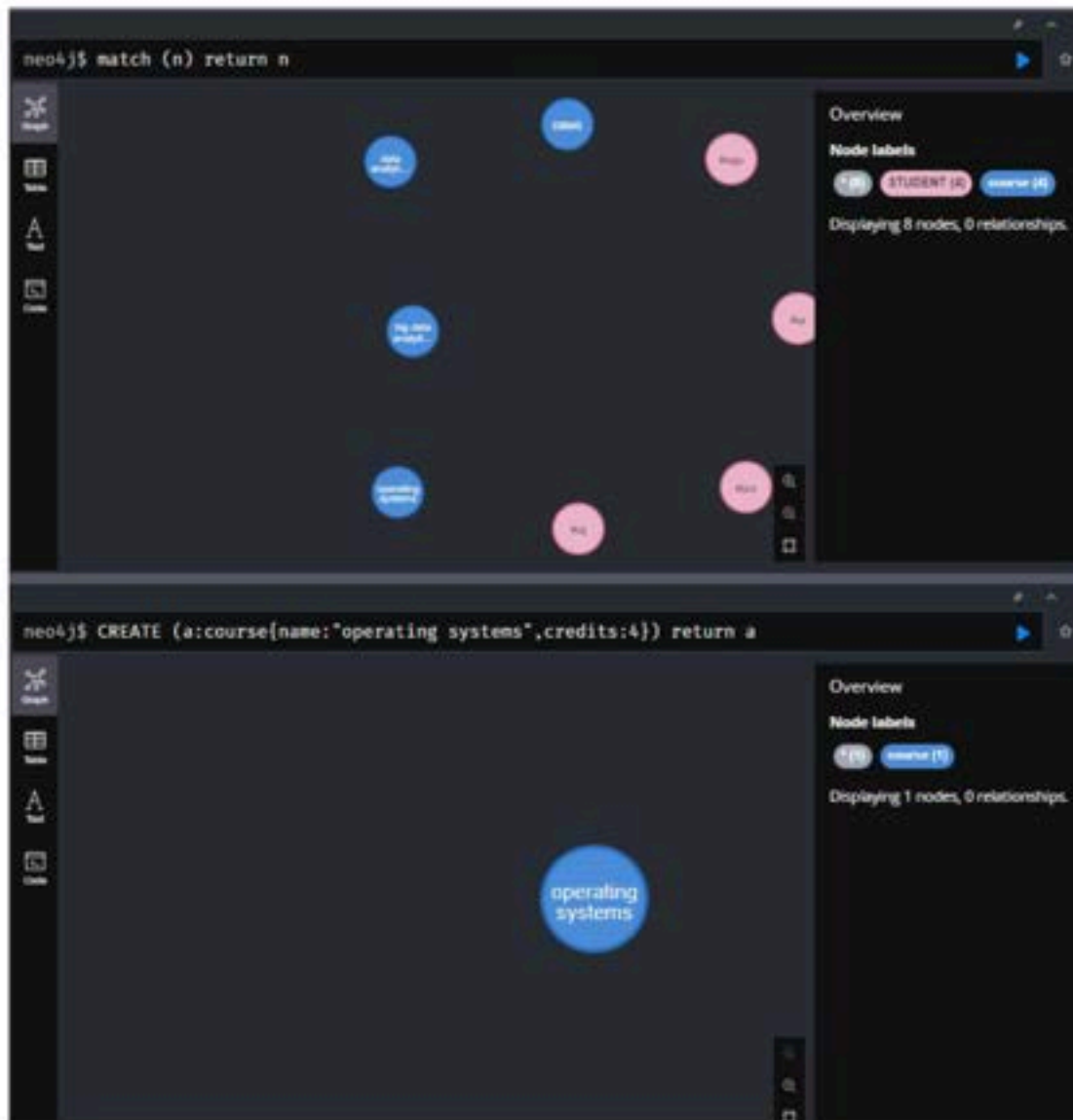
```
neo4j> CREATE (a: STUDENT {name:"Ra0",id:14})
```

Added 1 node, created 1 node, and 0 properties, completed after 1 ms.

```
neo4j> CREATE (a: STUDENT {name:"Rav1",id:13})
```

Added 1 node, created 1 node, and 0 properties, completed after 1 ms.


```
1 MATCH (a:TEACHER), (b:COURSE) Where a.name = "Prof. Aman" AND b.name = "DBMS" OR b.name
2 CREATE (a)-[r:TEACHES]->(b)
3
```



12/3/25 Lab 3

noisyf commands

noisyf \$ create (a: STUDENT {name: "Raj", id: 123})

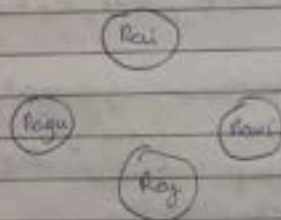
Output

Added 1 label created 1 node set a property completed after 16 ms

> Gremlin (a: STUDENT {name: "Raj", id: 123})

> Gremlin (a: STUDENT {name: "Raj", id: 143})

noisyf \$ MATCH(a) Return n

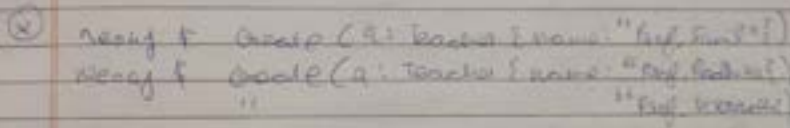


noisyf \$ create (a: course {name: "DBMS", credits: 4})

create (a: course {name: "data analytics", credits: 4}) return

create (a: course {name: "operating system", credits: 4})

create (a: course {name: "big data analytics", credits: 4})



```

(8)
def f(a: String, b: String):
    a = "hai"
    b = "hai"
    print(a, b)
    return a, b

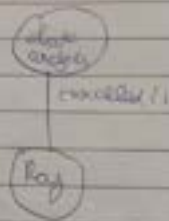
```

Spencer for Nash

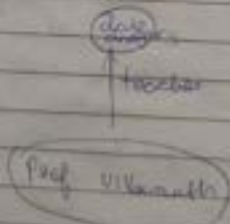


① for teacher

$\text{match } (a: \text{STUDENT}), (b: \text{course})$
 where $a.\text{name} = \text{"Ragu"}$ AND $b.\text{name}$
 $= \text{"data analytics"}$ $\text{match } (a) -$
 $[\exists x: \text{Enrolled In}] \rightarrow (b)$
 return a, b

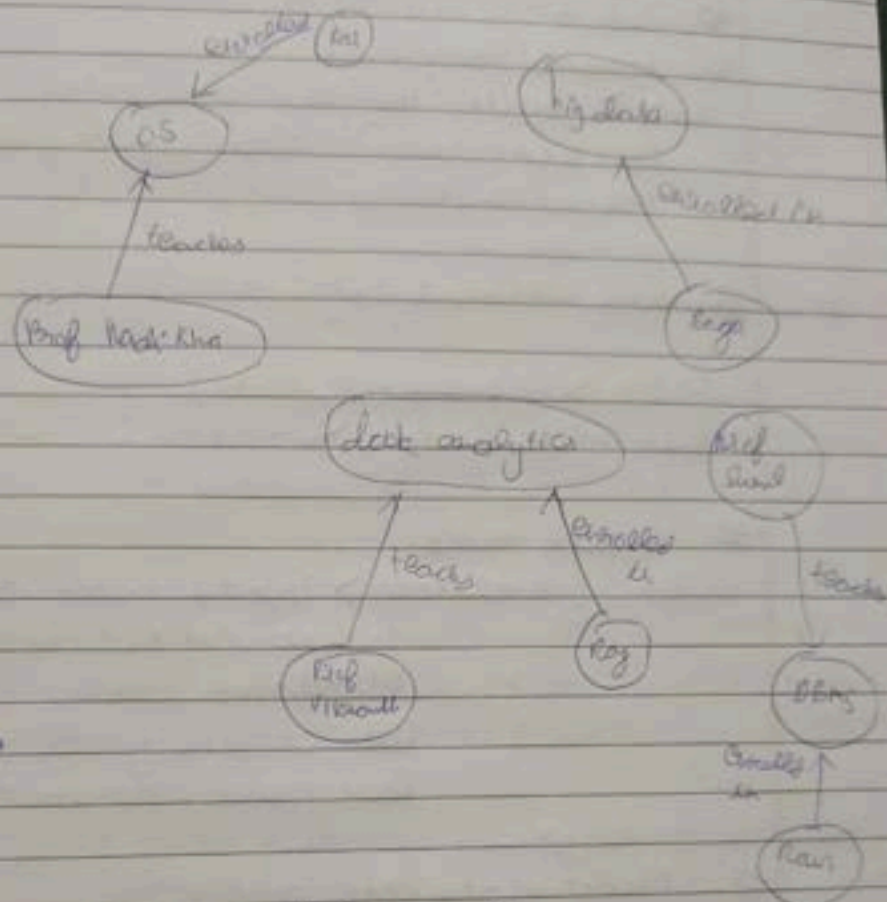


② $\text{match } (a: \text{TEACHER}), (b: \text{course})$
 where $a.\text{name} = \text{"Prof. Vignarath"}$ AND
 $b.\text{name} = \text{"data analytics"}$
 $\text{match } (a) - [\exists x: \text{teacher}] \rightarrow (b)$
 return a, b



store
67

query 5 match (a) return a



④ Match (a:teacher) (b:course) where a.name = "Prof. Mittal" AND (b.name = "DBMS" OR b.name = "DS")

(create(a) - [r:teacher]) → (b)

⑤ Match (a:Student & name: "RAJ") - [r] (b)
DELETE r

Lab 4: Cassandra

Question: 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

Code with Output:

1BM22CS158_BDA_Lab4_cassandra

1. Create Keyspace
2. Use the Keyspace
3. Create Table
4. Inserting Data into Table
5. View Table Data

```
bmscscie@bmscscie-HP-Elite-Tower-800-G9-Desktop-PC:~$ cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.1.0 | Cassandra 4.1.4 | CQL spec 3.4.0 | Native protocol v5]
Use HELP for help.
cqlsh> CREATE KEYSPACE Students WITH REPLICATION = {'class': 'SimpleStrategy', 'replication_factor': 1};
AlreadyExists: Keyspace 'Students' already exists
cqlsh> CREATE KEYSPACE school_data WITH REPLICATION = {'class': 'SimpleStrategy', 'replication_factor': 1};
cqlsh> USE school_data;
cqlsh:school_data>
cqlsh:school_data> CREATE TABLE student_info (
...     Roll_No int PRIMARY KEY,
...     StudName text,
...     DateofJoining timestamp,
...     last_exam_percent double
... );
cqlsh:school_data> INSERT INTO student_info (Roll_No, StudName, DateofJoining, last_exam_percent)
... VALUES (1, 'Asha', '2012-03-12', 79.9);
cqlsh:school_data> INSERT INTO student_info (Roll_No, StudName, DateofJoining, last_exam_percent)
... VALUES (2, 'Kiran', '2012-03-12', 89.9);
cqlsh:school_data> INSERT INTO student_info (Roll_No, StudName, DateofJoining, last_exam_percent)
... VALUES (3, 'Tarun', '2012-03-12', 78.8);
cqlsh:school_data> SELECT * FROM student_info;

roll_no | dateofjoining | last_exam_percent | studname
-----|-----|-----|-----
1 | 2012-03-12 18:30:00.000000+0000 | 79.9 | Asha
2 | 2012-03-12 18:30:00.000000+0000 | 89.9 | Kiran
3 | 2012-03-12 18:30:00.000000+0000 | 78.8 | Tarun

(3 rows)
cqlsh:school_data>
```

6. Create Index
7. Select Data Using Index

```
bmscscie@bmscscie-HP-Elite-Tower-800-G9-Desktop-PC:~$ cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.1.0 | Cassandra 4.1.4 | CQL spec 3.4.0 | Native protocol v5]
Use HELP for help.
cqlsh> USE school_data;
cqlsh:school_data> CREATE INDEX ON student_info (StudName);
cqlsh:school_data> SELECT * FROM student_info WHERE StudName = 'Asha';

roll_no | dateofjoining | last_exam_percent | studname
-----|-----|-----|-----
1 | 2012-03-12 18:30:00.000000+0000 | 79.9 | Asha

(1 rows)
cqlsh:school_data> UPDATE student_info SET StudName = 'David Sheen' WHERE Roll_No = 2;
```

8. Update Data

9. Delete Data

10. Add a Set or List Collection

11. Update Collections

```

Innocent@Innocent-HP-Elite-Tower-800-G3-Desktop-PC:~$ cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 4.1.0 | Cassandra 4.1.4 | CQL spec 3.4.6 | Native protocol v5]
Use HELP for help.
cqlsh> USE school_data;
cqlsh:school_data> SELECT * FROM student_info;

roll_no | dateofjoining | last_exam_percent | studname
-----|-----|-----|-----
1 | 2012-03-11 18:38:00.000000+0000 | 79.9 | Asha
2 | 2012-03-11 18:38:00.000000+0000 | 83.9 | David Sheen
3 | 2012-03-11 18:38:00.000000+0000 | 78 | Tarun

(3 rows)
cqlsh:school_data> DELETE FROM student_info WHERE Roll_No = 2;
cqlsh:school_data> SELECT * FROM student_info;

roll_no | dateofjoining | last_exam_percent | studname
-----|-----|-----|-----
1 | 2012-03-11 18:38:00.000000+0000 | 79.9 | Asha
3 | 2012-03-11 18:38:00.000000+0000 | 78 | Tarun

(2 rows)
cqlsh:school_data> ALTER TABLE student_info ADD hobbies SET<text>;
cqlsh:school_data> ALTER TABLE student_info ADD languages LIST<text>;
cqlsh:school_data> UPDATE student_info SET hobbies = hobbies + {'Chess', 'Table Tennis'} WHERE Roll_No = 1;
cqlsh:school_data> UPDATE student_info SET languages = languages + {'English', 'French'} WHERE Roll_No = 1;
cqlsh:school_data> UPDATE student_info SET languages = languages || ['English', 'French'] WHERE Roll_No = 1;
UPDATE syntax at char 47
UPDATE student_info SET languages = languages || ['English', 'French'] WHERE Roll_No = 1;
cqlsh:school_data> SELECT * FROM student_info;

roll_no | dateofjoining | hobbies | languages | last_exam_percent | studname
-----|-----|-----|-----|-----|-----
1 | 2012-03-11 18:38:00.000000+0000 | {'Chess', 'Table Tennis'} | null | 79.9 | Asha
3 | 2012-03-11 18:38:00.000000+0000 | null | null | 78 | Tarun

(2 rows)
cqlsh:school_data>

```

```

(2 rows)
cqlsh:school_data> ALTER TABLE student_info ADD hobbies SET<text>;
cqlsh:school_data> ALTER TABLE student_info ADD languages LIST<text>;
cqlsh:school_data> UPDATE student_info SET hobbies = hobbies + {'Chess', 'Table Tennis'} WHERE Roll_No = 1;
cqlsh:school_data> UPDATE student_info SET languages = languages + {'English', 'French'} WHERE Roll_No = 1;
cqlsh:school_data> UPDATE student_info SET languages = languages || ['English', 'French'] WHERE Roll_No = 1;
UPDATE syntax at char 47
UPDATE student_info SET languages = languages || ['English', 'French'] WHERE Roll_No = 1;
cqlsh:school_data> SELECT * FROM student_info;

roll_no | dateofjoining | hobbies | languages | last_exam_percent | studname
-----|-----|-----|-----|-----|-----
1 | 2012-03-11 18:38:00.000000+0000 | {'Chess', 'Table Tennis'} | null | 79.9 | Asha
3 | 2012-03-11 18:38:00.000000+0000 | null | null | 78 | Tarun

(2 rows)
cqlsh:school_data> UPDATE student_info SET languages = languages + ['English', 'French'] WHERE Roll_No = 1;
cqlsh:school_data> SELECT * FROM student_info;

roll_no | dateofjoining | hobbies | languages | last_exam_percent | studname
-----|-----|-----|-----|-----|-----
1 | 2012-03-11 18:38:00.000000+0000 | {'Chess', 'Table Tennis'} | ['English', 'French'] | 79.9 | Asha
3 | 2012-03-11 18:38:00.000000+0000 | null | null | 78 | Tarun

(2 rows)
cqlsh:school_data>

```

14. Batch Insert Operations

```
cqlsh:school_data> BEGIN BATCH
... INSERT INTO student_info (roll_no, studname, dateofjoining, last_exam_percent) VALUES (4, 'Saurth', '2012-03-12', 90.9);
... INSERT INTO student_info (roll_no, studname, dateofjoining, last_exam_percent) VALUES (5, 'Smitha', '2012-03-12', 67.9);
... INSERT INTO student_info (roll_no, studname, dateofjoining, last_exam_percent) VALUES (6, 'Neben', '2012-03-12', 56.9);
... APPLY BATCH;

cqlsh:school_data> SELECT * FROM student_info;
```

roll_no	dateofjoining	studname	hobbies	languages	last_exam_percent	studname
5	2012-03-12 18:30:00.000000+0000		null	null	67.9	Smitha
4	2012-03-12 18:30:00.000000+0000		'Chess', 'Table Tennis'	'English', 'French'	90.9	Saurth
6	2012-03-12 18:30:00.000000+0000		null	null	56.9	Neben
3	2012-03-12 18:30:00.000000+0000		null	null	78	Tarun

(5 rows)
cqlsh:school_data>

13. Describe Keyspaces and Tables

```
cqlsh:school_data> DESCRIBE KEYSPACES;
```

keyspace	newstudents	students	students9	system_schema
education	school_data	students1	system	system_traces
employee	stud	students3	system_auth	system_views
newstudent	student	students8	system_distributed	system_virtual_schema

```
cqlsh:school_data> DESCRIBE TABLES;
```

```
student_info
```

```
cqlsh:school_data> DESCRIBE TABLE student_info;
```

```
CREATE TABLE school_data.student_info (
  roll_no int PRIMARY KEY,
  dateofjoining timestamp,
  last_exam_percent double,
  studname text,
  hobbies set<text>,
  languages list<text>
) WITH additional_write_policy = '99p'
AND bloom_filter_fp_chance = 0.01
AND caching = {'keys': 'ALL', 'rows_per_partition': 'NONE'}
AND idc = false
AND comment = ''
AND compaction = {'class': 'org.apache.cassandra.db.compaction.SizeTieredCompactionStrategy', 'max_threshold': '32', 'min_threshold': '4'}
AND compression = {'chunk_length_in_kb': '16', 'class': 'org.apache.cassandra.io.compress.LZ4Compressor'}
AND crc_table = 'default'
AND crc_check_chance = 1.0
AND default_time_to_live = 0
AND extensions = {}
AND gc_grace_seconds = 8640000
AND max_index_interval = 2048
AND memtable_flush_period_in_ms = 0
AND min_index_interval = 128
AND read_repair = 'BLOCKING'
AND speculative_retry = '99p';
```

```
CREATE INDEX student_info_studname_idx ON school_data.student_info (studname);
```

```
cqlsh:school_data> SELECT * FROM student_info;
```

roll_no	dateofjoining	studname	hobbies	languages	last_exam_percent	studname
5	2012-03-12 18:30:00.000000+0000		null	null	67.9	Smitha
4	2012-03-12 18:30:00.000000+0000		'Chess', 'Table Tennis'	'English', 'French'	90.9	Saurth
6	2012-03-12 18:30:00.000000+0000		null	null	56.9	Neben
3	2012-03-12 18:30:00.000000+0000		null	null	78	Tarun

(5 rows)
cqlsh:school_data>

2. Exporting and Importing Data

Export to CSV:

Import from CSV:

```
(5 rows)
cqlsh:school_data> COPY student_info (roll_no, studname, dateofjoining, last_exam_percent) TO 'student_info_export.csv';
Using 10 child processes

Starting copy of school_data.student_info with columns (roll_no, studname, dateofjoining, last_exam_percent).
Processed 5 rows: Rate: 16 rows/s; Avg. rate: 16 rows/s
3 rows exported to 1 files in 0.140 seconds.

cqlsh:school_data> COPY student_info (roll_no, studname, dateofjoining, last_exam_percent) FROM 'student_info_import.csv';
Using 10 child processes

Starting copy of school_data.student_info with columns (roll_no, studname, dateofjoining, last_exam_percent).
Using 10 rows: 0 rows imported; 10 rows not imported; 0 rows not imported; 0 rows not imported; 0 rows not imported; 0 rows not imported; 0 rows not imported; 0 rows not imported; 0 rows not imported; 0 rows not imported.
Processed 0 rows: Rate: 0 rows/s; Avg. rate: 0 rows/s
0 rows imported from 0 files in 0.007 seconds (0 skipped).
cqlsh:school_data>
```


11/4/25

lastandra

cqlsh

create keyspace

```
cqlsh > Create KeySPACE school_data WITH  
REPLICATION = { 'class' : 'SimpleStrategy',  
'replication_factor' : 1 };
```

use student_data

```
cqlsh > USE school_data;  
cqlsh : school_data >
```

create table

```
cqlsh : school_data > Create Table student_info  
(roll_no int PRIMARY KEY,  
student_name text,  
date_of_joining timestamp,  
last_exam_percent double)
```

insert data into Table

```
cqlsh : school_data > Insert Into student_info  
(roll_no, student_name, date_of_joining, last_exam_percent) values (1, 'Ashu', '2012-02-12', 79.9);
```

```
cqlsh : school_data > Insert Into student_info  
(roll_no, student_name, date_of_joining, last_exam_percent) values (2, 'Karan', '2012-03-11', 78);
```

view table

sql> select * from student_info

roll no	date of joining	last exam percent	student name
1	2012-02-11 18:10	79.9	Ashu
2	2012-03-11 18:10	78	Karan
3	2012-03-11 18:10	78	Karan

store
67

Create Index

sql> school_data > CREATE INDEX student_info
(StudentInfo),

Coln: school_data >

→

Selects data using Index

sql> school_data > SELECT * FROM student_info
WHERE StudentInfo = 'Ashu';

Roll no	Date of joining	Percentage	Student
1	2012-03-11 18:30	79.9	Ashu

Delete Data

Coln: school_data > DELETE FROM student_info WHERE

Coln: school_data > SELECT * FROM student_info;

Roll no	Date of joining	Percentage	Student
1	2012-03-11	79.9	Ashu
3	2012-03-11	78	Amu

Add a Set of a list Collection

sql> school_data > ALTER TABLE student_info
ADD hobbies SET <text>

sql> school_data > ALTER TABLE student_info
ADD long_name LIST <text>

Update

sql> school_data > UPDATE student_info SET
Hobbies = hobbies + { 'chess', 'Table Tennis' }
WHERE Roll no = 1;

Adapt select * from Student Info

Roll no	date of joining	hobbies	Language	Mark
1	2012-02-11	den IT	null	78
2	2012-03-11	null	null	78

Update Language

sql> school-data > update studentinfo
 SET Language = 'English'
 FROM studentinfo WHERE Roll no = 1,

Roll no	date of join	hobbies	Language	Mark
1	2012-02-11	den IT	English	78
2	2012-03-11	null	null	78

Batch Insert

sql> school-data > begin Batch
 insert into studentinfo (Roll no, date of join, Mark)
 values (4, 'Saurabh', '2012-03-11', 90),
 (5, 'Smitha', '2012-03-11', 56),
 (6, 'Rohan', '2012-03-11', 78);

Describe Vagpavkar

sql> school-data > describe Vagpavkar,
 Column Name Data Type
 Education School data
 Country en
 Address Mumbai

Decoded Index
after school data > Decoded Index,
Student Info

Decoded Specific Index
after school data > Decoded Index, Student Info

Export to CSV
after school data > copy student info (Roll no, student name, Date of Birth, sex, etc.) to 'student info export.csv',
use it child processes

q/p Show copy of school data, student info
with columns [roll no, student name, dob, sex]

Processed 5 rows rate 50 rows/5 sec rate
15 rows/5

5 rows exported to 'file' in 0.142 sec

Export to CSV

Showing copy of School data, Student info
with columns.

8/14

Lab 5: Cassandra

Question: Perform the following DB operations using Cassandra.

1. Create a keyspace by name Employee
2. Create a column family by name Employee-Info with attributes Emp_Id Primary Key, Emp_Name, Designation, Date_of_Joining, Salary, Dept_Name
3. Insert the values into the table in batch
4. Update Employee name and Department of Emp-Id 121
5. Sort the details of Employee records based on salary
6. Alter the schema of the table Employee_Info to add a column Projects which stores a set of Projects done by the corresponding Employee.
7. Update the altered table to add project names.
8. Create a TTL of 15 seconds to display the values of Employees.

Code with Output:

1BM22CS158_Lab-5_BDA_cassandra

Part 1: Employee Database (Table name changed to **employee_details**)

#Create Keyspace

#Create Table **employee_details**

#Insert Values in Batch

```
bmscscse@bmscscse-HP-Elite-Tower-890-C9-Desktop-PC: -  
bmscscse@bmscscse-HP-Elite-Tower-890-C9-Desktop-PC:~$ cqlsh  
Connected to Test Cluster at 127.0.0.1:9042  
[cqlsh 6.1.0 | Cassandra 4.1.4 | CQL spec 3.4.6 | Native protocol v5]  
Use HELP for help.  
cqlsh> CREATE KEYSPACE employee_db WITH replication = {'class': 'SimpleStrategy', 'replication_factor': 1};  
AlreadyExists: Keyspace 'employee_db' already exists  
cqlsh> USE employee_db;  
cqlsh:employee_db> CREATE TABLE employee_details (  
...     emp_id int PRIMARY KEY,  
...     emp_name text,  
...     designation text,  
...     date_of_joining date,  
...     salary decimal,  
...     dept_name text  
... );  
cqlsh:employee_db> BEGIN BATCH  
... INSERT INTO employee_details (emp_id, emp_name, designation, date_of_joining, salary, dept_name)  
... VALUES (121, 'Alice', 'Manager', '2018-05-20', 75000.00, 'HR');  
... INSERT INTO employee_details (emp_id, emp_name, designation, date_of_joining, salary, dept_name)  
... VALUES (122, 'Bob', 'Developer', '2020-07-01', 60000.00, 'IT');  
... INSERT INTO employee_details (emp_id, emp_name, designation, date_of_joining, salary, dept_name)  
... VALUES (123, 'Charlie', 'Analyst', '2021-09-10', 50000.00, 'Finance');  
... APPLY BATCH;  
cqlsh:employee_db> select * from employee_details  
... ;  
  
emp_id | date_of_joining | dept_name | designation | emp_name | salary  
-----  
123 | 2021-09-10 | Finance | Analyst | Charlie | 50000.00  
122 | 2020-07-01 | IT | Developer | Bob | 60000.00  
121 | 2018-05-20 | HR | Manager | Alice | 75000.00  
(3 rows)  
cqlsh:employee_db> |
```

```
#Update emp_name and dept_name of emp_id 121
#Sort Employees by Salary
#Alter Table to Add projects Column
#Update Projects for Employee
```

```
cqlsh:employee_db> UPDATE employee_details
... SET emp_name = 'Alicia', dept_name = 'Operations'
... WHERE emp_id = 121;
cqlsh:employee_db> select * from employee_details ;
```

emp_id	date_of_joining	dept_name	designation	emp_name	salary
123	2021-09-10	Finance	Analyst	Charlie	50000.00
122	2020-07-01	IT	Developer	Bob	60000.00
121	2018-05-20	Operations	Manager	Alicia	75000.00

```
(3 rows)
cqlsh:employee_db> -- This needs ALLOW FILTERING as sorting isn't supported directly:
cqlsh:employee_db> SELECT * FROM employee_details WHERE salary > 0 ALLOW FILTERING;
```

emp_id	date_of_joining	dept_name	designation	emp_name	salary
123	2021-09-10	Finance	Analyst	Charlie	50000.00
122	2020-07-01	IT	Developer	Bob	60000.00
121	2018-05-20	Operations	Manager	Alicia	75000.00

```
(3 rows)
cqlsh:employee_db> ALTER TABLE employee_details ADD projects set<text>;
```

```
#Update Projects for Employee
#Insert with TTL of 15 Seconds
```

```

sqlsh> employee_db> UPDATE employee_details
... SET projects = ('Project A', 'Project B')
... WHERE emp_id = 123;
sqlsh> employee_db> INSERT INTO employee_details (emp_id, emp_name, designation, date_of_joining, salary, dept_name)
... VALUES (124, 'Eve', 'Intern', '2023-01-15', 30000.00, 'Marketing') USING 770 13;
sqlsh> employee_db> select * from employee_details ;

emp_id | date_of_joining | dept_name | designation | emp_name | projects | salary
-----|-----|-----|-----|-----|-----|-----
123 | 2023-09-10 | Finance | Analyst | Charlie | null | 50000.00
122 | 2020-07-01 | IT | Developer | Bob | null | 40000.00
121 | 2020-01-20 | Operations | Manager | Alice | ('Project A', 'Project B') | 70000.00
(3 rows)

sqlsh> employee_db> INSERT INTO employee_details (emp_id, emp_name, designation, date_of_joining, salary, dept_name) VALUES (124, 'Eve', 'Intern', '2023-01-15', 30000.00, 'Marketing')
sqlsh> employee_db> select * from employee_details ;

emp_id | date_of_joining | dept_name | designation | emp_name | projects | salary
-----|-----|-----|-----|-----|-----|-----
123 | 2023-09-10 | Finance | Analyst | Charlie | null | 50000.00
122 | 2020-07-01 | IT | Developer | Bob | null | 40000.00
121 | 2020-01-20 | Operations | Manager | Alice | ('Project A', 'Project B') | 70000.00
124 | 2023-01-15 | Marketing | Intern | Eve | null | 30000.00
(4 rows)

sqlsh> employee_db> select * from employee_details ;

emp_id | date_of_joining | dept_name | designation | emp_name | projects | salary
-----|-----|-----|-----|-----|-----|-----
123 | 2023-09-10 | Finance | Analyst | Charlie | null | 50000.00
122 | 2020-07-01 | IT | Developer | Bob | null | 40000.00
121 | 2020-01-20 | Operations | Manager | Alice | ('Project A', 'Project B') | 70000.00
124 | 2023-01-15 | Marketing | Intern | Eve | null | 30000.00
(4 rows)

sqlsh> employee_db> INSERT INTO employee_details (emp_id, emp_name, designation, date_of_joining, salary, dept_name) VALUES (124, 'Eve', 'Intern', '2023-01-15', 30000.00, 'Marketing') USING 770 13;
sqlsh> employee_db> select * from employee_details ;

emp_id | date_of_joining | dept_name | designation | emp_name | projects | salary
-----|-----|-----|-----|-----|-----|-----
123 | 2023-09-10 | Finance | Analyst | Charlie | null | 50000.00
122 | 2020-07-01 | IT | Developer | Bob | null | 40000.00
121 | 2020-01-20 | Operations | Manager | Alice | ('Project A', 'Project B') | 70000.00
124 | 2023-01-15 | Marketing | Intern | Eve | null | 30000.00
(4 rows)

sqlsh> employee_db> select * from employee_details ;

emp_id | date_of_joining | dept_name | designation | emp_name | projects | salary
-----|-----|-----|-----|-----|-----|-----
123 | 2023-09-10 | Finance | Analyst | Charlie | null | 50000.00
122 | 2020-07-01 | IT | Developer | Bob | null | 40000.00
121 | 2020-01-20 | Operations | Manager | Alice | ('Project A', 'Project B') | 70000.00
124 | 2023-01-15 | Marketing | Intern | Eve | null | 30000.00
(4 rows)

sqlsh> employee_db> select * from employee_details ;

emp_id | date_of_joining | dept_name | designation | emp_name | projects | salary
-----|-----|-----|-----|-----|-----|-----
123 | 2023-09-10 | Finance | Analyst | Charlie | null | 50000.00
122 | 2020-07-01 | IT | Developer | Bob | null | 40000.00
121 | 2020-01-20 | Operations | Manager | Alice | ('Project A', 'Project B') | 70000.00
124 | 2023-01-15 | Marketing | Intern | Eve | null | 30000.00
(4 rows)

```

Part 2: Library Database (New Table Names)

Create Keyspace

Create Tables

A. Table: `library_student_info`

B. Table: `book_counter_info`

3 Insert Data in Batch

```

$ cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.1.0 | Cassandra 4.1.4 | CQL spec 3.4.4 | Native protocol v5]
Use HELP for help.
cqlsh> CREATE KEYSPACE library_db WITH replication = {'class': 'SimpleStrategy', 'replication_factor': 1};
cqlsh> USE library_db;
cqlsh:library_db> CREATE TABLE library_student_info (
...     stud_id int PRIMARY KEY,
...     stud_name text,
...     book_name text,
...     book_id int,
...     date_of_issue date
... );
cqlsh:library_db> CREATE TABLE book_counter_info (
...     stud_id int,
...     book_name text,
...     counter_value counter,
...     PRIMARY KEY (stud_id, book_name)
... );
cqlsh:library_db> BEGIN BATCH
... INSERT INTO library_student_info (stud_id, stud_name, book_name, book_id, date_of_issue)
... VALUES (112, 'David', 'BDA', 401, '2024-03-12');
... UPDATE book_counter_info SET counter_value = counter_value + 1
... WHERE stud_id = 112 AND book_name = 'BDA';
... APPLY BATCH;
SerialIdRequest: Error from server: code=2000 [Invalid query] message="counter and non-counter mutations cannot exist in the same batch"

```

(No Mixing Counters in Batch):

You can repeat the **UPDATE** if you want to increment the counter multiple times.

To Simulate Borrowing Book "BDA" 2 Times by Student 112

Display Table & Increase Counter

Query: Student 112 took "BDA" 2 times

```

cqlsh:library_db> -- First: Insert normal data (non-counter)
cqlsh:library_db> INSERT INTO library_student_info (stud_id, stud_name, book_name, book_id, date_of_issue)
... VALUES (112, 'David', 'BDA', 401, '2024-03-12');
cqlsh:library_db> -- Then: Update the counter table separately
cqlsh:library_db> UPDATE book_counter_info
... SET counter_value = counter_value + 1
... WHERE stud_id = 112 AND book_name = 'BDA';
cqlsh:library_db> -- Insert once (already done above)
cqlsh:library_db> -- Increment counter again
cqlsh:library_db> UPDATE book_counter_info
... SET counter_value = counter_value + 1
... WHERE stud_id = 112 AND book_name = 'BDA';
cqlsh:library_db> SELECT * FROM library_student_info;

stud_id | book_id | book_name | date_of_issue | stud_name
-----
112 | 401 | BDA | 2024-03-12 | David

(1 rows)
cqlsh:library_db> SELECT * FROM book_counter_info;

stud_id | book_name | counter_value
-----
112 | BDA | 2

(1 rows)
cqlsh:library_db>
cqlsh:library_db> -- Increment counter again:
cqlsh:library_db> UPDATE book_counter_info SET counter_value = counter_value + 1
... WHERE stud_id = 112 AND book_name = 'BDA';
cqlsh:library_db> SELECT counter_value FROM book_counter_info
... WHERE stud_id = 112 AND book_name = 'BDA';

counter_value
-----
3

```

Export Table to CSV

```

(1 rows)
cqlsh:library_db> cqlsh -e "COPY library_db.library_student_info TO 'library_export.csv' WITH HEADER = TRUE;"

```

```

cqlsh:library_db> cqlsh> COPY library_student_info (stud_id, stud_name, book_name, book_id, date_of_issue) FROM '/home/broceca/Documents/library_data.csv' WITH HEADER = TRUE;

```


Page 67

Employee Table

Part 1: Employee Table Creation

Create Table

```

CREATE TABLE Employee (
    empid INT PRIMARY KEY,
    empname VARCHAR(50),
    dept VARCHAR(20),
    salary DECIMAL(10,2),
    hiredate DATE
);
        
```

Insert Data

```

INSERT INTO Employee (empid, empname, dept, salary, hiredate)
VALUES (101, 'John', 'Sales', 1200, '2013-01-01'),
       (102, 'Jane', 'IT', 1500, '2013-02-01'),
       (103, 'Mike', 'Marketing', 1800, '2013-03-01');
        
```

Update Data

```

UPDATE Employee SET salary = salary * 1.1;
        
```

Delete Data

```

DELETE FROM Employee WHERE empid = 101;
        
```

Page 68

Employee Table

Part 2: Employee Table

Create Table

```

CREATE TABLE Employee (
    empid INT PRIMARY KEY,
    empname VARCHAR(50),
    dept VARCHAR(20),
    salary DECIMAL(10,2),
    hiredate DATE
);
        
```

Insert Data

```

INSERT INTO Employee (empid, empname, dept, salary, hiredate)
VALUES (101, 'John', 'Sales', 1200, '2013-01-01'),
       (102, 'Jane', 'IT', 1500, '2013-02-01'),
       (103, 'Mike', 'Marketing', 1800, '2013-03-01');
        
```

Update Data

```

UPDATE Employee SET salary = salary * 1.1;
        
```

Delete Data

```

DELETE FROM Employee WHERE empid = 101;
        
```

Page 69

Employee Table

Part 3: Employee Table

Create Table

```

CREATE TABLE Employee (
    empid INT PRIMARY KEY,
    empname VARCHAR(50),
    dept VARCHAR(20),
    salary DECIMAL(10,2),
    hiredate DATE
);
        
```

Insert Data

```

INSERT INTO Employee (empid, empname, dept, salary, hiredate)
VALUES (101, 'John', 'Sales', 1200, '2013-01-01'),
       (102, 'Jane', 'IT', 1500, '2013-02-01'),
       (103, 'Mike', 'Marketing', 1800, '2013-03-01');
        
```

Update Data

```

UPDATE Employee SET salary = salary * 1.1;
        
```

Delete Data

```

DELETE FROM Employee WHERE empid = 101;
        
```

Page 70

Employee Table

Part 4: Employee Table

Create Table

```

CREATE TABLE Employee (
    empid INT PRIMARY KEY,
    empname VARCHAR(50),
    dept VARCHAR(20),
    salary DECIMAL(10,2),
    hiredate DATE
);
        
```

Insert Data

```

INSERT INTO Employee (empid, empname, dept, salary, hiredate)
VALUES (101, 'John', 'Sales', 1200, '2013-01-01'),
       (102, 'Jane', 'IT', 1500, '2013-02-01'),
       (103, 'Mike', 'Marketing', 1800, '2013-03-01');
        
```

Update Data

```

UPDATE Employee SET salary = salary * 1.1;
        
```

Delete Data

```

DELETE FROM Employee WHERE empid = 101;
        
```


Lab 6 :

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

Code with Output:

1BM22CS158_Lab6_Hadoop
Start-all.sh

```
hadoop@bnscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ start-all.sh
WARNING: Attempting to start all Hadoop daemons on hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenode.sh [localhost]
localhost: namenode is running as process 1786. Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanode.sh
localhost: datanode is running as process 1880. Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondarynamenode.sh [localhost: bnscecse-HP-Elite-Tower-800-G9-Desktop-PC]
Starting resourcemanager.sh [localhost: bnscecse-HP-Elite-Tower-800-G9-Desktop-PC]
localhost: secondarynamenode is running as process 1814. Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
Starting resourcemanager.sh
localhost: resourcemanager is running as process 1814. Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting hadoopmanager.sh
localhost: hadoopmanager is running as process 1814. Stop it first and ensure /tmp/hadoop-hadoop-hadoopmanager.pid file is empty before retry.
hadoop@bnscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$
```

1. Switch to hduser

su hduser right user (hadoop), so no need to su hduser.

#jps

```
hadoop@bnscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ jps
9877 Jps
6646 NodeManager
5895 DataNode
6188 SecondaryNameNode
5789 NameNode
6478 ResourceManager
hadoop@bnscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$
```

#which hadoop

```
hadoop@bnscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ which hadoop
/home/hadoop/hadoop/bin/hadoop
```

```

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ which hadoop
/home/hadoop/hadoop/bin/hadoop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cd /home/hadoop/hadoop/bin/hadoop/sbin
./start-dfs.sh
./start-yarn.sh
bask: cd: /home/hadoop/hadoop/bin/hadoop/sbin: not a directory
bask: ./start-dfs.sh: No such file or directory
bask: ./start-yarn.sh: No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cd /home/hadoop/hadoop/bin/hadoop/sbin
bask: cd: /home/hadoop/hadoop/bin/hadoop/sbin: not a directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cd /home/hadoop/hadoop/bin/hadoop
bask: cd: /home/hadoop/hadoop/bin/hadoop: not a directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cd hadoop
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop$ ls
bin  etc  include  lib  libexec  binary  common-binary  LICENSE.txt  logs  NOTICE-binary  NOTICE.txt  README.txt  share  share-docs
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop$ cd src
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/src$ ./start-dfs.sh
Starting namenode on [localhost]
localhost: namenode is running as process 1789. Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 1865. Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [process-HP-Elite-Tower-800-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: secondarynamenode is running as process 4188. Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/src$ ./start-yarn.sh
Starting resourcemanager
resourcemanager is running as process 4476. Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 5646. Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/src$

```

Create and Upload the File

Directory change and file creation hadoop file

```

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop$ hadoop fs -ls /
Found 4 items
drwxr-xr-x - hadoop supergroup 0 2024-05-14 14:55 /HELLO
drwxr-xr-x - hadoop supergroup 0 2024-05-14 14:44 /abc
drwxr-xr-x - hadoop supergroup 0 2024-05-21 15:09 /output
drwxr-xr-x - hadoop supergroup 0 2024-05-21 14:59 /rgs
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop$ hadoop fs -mkdir /shu158
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop$ hadoop fs -copyfromlocal /home/hadoop/Desktop/hadoop158.txt /shu158/hadoop158.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop$ hadoop fs -copyfromlocal /home/hadoop/Desktop/hadoop158.txt /shu158/hadoop158.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop$

```

```

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/hadoop/sbin$ cd -
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cd desktop/
bash: cd: desktop/: No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cd Desktop/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ nano hadoop158
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hadoop fs -ls
ls: '.': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hadoop fs -ls /
Found 4 items
drwxr-xr-x - hadoop supergroup 0 2024-05-14 14:55 /HELLO
drwxr-xr-x - hadoop supergroup 0 2024-05-14 14:44 /abc
drwxr-xr-x - hadoop supergroup 0 2024-05-21 15:09 /output
drwxr-xr-x - hadoop supergroup 0 2024-05-21 14:59 /rgs

```


11/11/17

Handicap

creating jobs & H.C.
 Hand - out - sh
 O/P: what was outcome of this

J.F.S.
 TRS

O/P:
 Standard
 Data book
 International Standard
 International
 International

with backup
 & with backup
 /hand/ backup/ hand/ win/ handicap

change directory & file location
 - \$ cd backup
 - /handicap & sh
 sh
 - /handicap & cd sh

check if you're still in the same
 & /hand - up sh
 & /hand - up sh

file location
 & make backup file
 & make a backup of project

Reference
 cd /handicap
 and make a backup
 & sh

make
 & make - classpath /handicap
 -d class /handicap

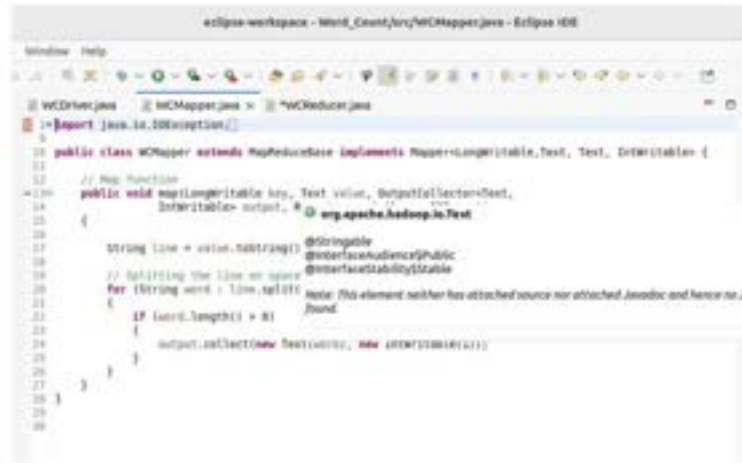
Java file

\$ Java -cub WordCount.java -c classes

added new file
 adding

1/1/17

Eclipse - jar files creation (WCMapper)



```

1 // WCMapper.java
2
3 import java.io.IOException;
4
5 public class WCMapper extends MapReduceBase implements Mapper<LongWritable, Text, Text, Text> {
6
7     // Map function
8     public void map(LongWritable key, Text value, OutputCollector<Text>
9         output, Reporter r) throws IOException {
10
11         String line = value.toString();
12
13         // Splitting the line on space
14         for (String word : line.split(" ")) {
15             // Note: This element neither has attached source nor attached javadoc and hence not found
16             output.collect(new Text(word), new LongWritable(1));
17         }
18     }
19 }

```

```

hadoop@bmsccse-HP-Elite-Tower-800-G9-Desktop-PC:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [bmsccse-HP-Elite-Tower-800-G9-Desktop-PC]
Starting resourcemanager
Starting nodemanagers
hadoop@bmsccse-HP-Elite-Tower-800-G9-Desktop-PC:~$ su huser
su: user huser does not exist or the user entry does not contain all the required fields
hadoop@bmsccse-HP-Elite-Tower-800-G9-Desktop-PC:~$ jps
7890 ResourceManager
8054 NodeManager
8480 Jps
7259 DataNode
5468 org.eclipse.equinox.launcher_1.6.1000.v20250227-1734.jar
7117 NameNode
7117 SecondaryNameNode

```

```

hadoop@bmsccse-HP-Elite-Tower-800-G9-Desktop-PC:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
LocalNode: namenode is running as process 7117. Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
LocalNode: datanode is running as process 7259. Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmsccse-HP-Elite-Tower-800-G9-Desktop-PC]
SecondaryNode: secondarynamenode is running as process 7117. Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
Starting resourcemanager
ResourceManager is running as process 7890. Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
LocalNode: nodemanager is running as process 8054. Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmsccse-HP-Elite-Tower-800-G9-Desktop-PC:~$ jps
7890 ResourceManager
8054 NodeManager
8480 Jps
7259 DataNode
5468 org.eclipse.equinox.launcher_1.6.1000.v20250227-1734.jar
7117 NameNode
7117 SecondaryNameNode

```

```

hadoop@bmsccse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -copyFromLocal /home/hadoop/working/Files.txt /user/hadoop/test.txt
hadoop@bmsccse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/working/hadoop.jar /home/hadoop/working/hadoop.jar WCMapper /user/hadoop/test.txt /user/hadoop/output
2025-04-25 15:11:00,470 INFO impl.MetricSystemImpl: Loaded properties from hadoop-metric2.properties
2025-04-25 15:11:00,474 INFO impl.MetricSystemImpl: Scheduled metric snapshot period at 10 second(s).
2025-04-25 15:11:00,474 INFO impl.MetricSystemImpl: JobTracker metric system started
2025-04-25 15:11:00,475 INFO impl.MetricSystemImpl: JobTracker metric system already initialized
2025-04-25 15:11:00,484 INFO mapreduce.JobResourceUploader: Hadoop command line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-04-25 15:11:00,490 INFO mapreduce.JobClient$ToolRunner: Total input files to process : 1
2025-04-25 15:11:00,491 INFO mapreduce.JobClient$ToolRunner: number of splits:1
2025-04-25 15:11:00,492 INFO mapreduce.JobClient$ToolRunner: Submitting the job. Job ID is job_1734100000000000000

```



```
hadoop@mscscse-MP-Elite-Tower-000-G9-Desktop-PC: ~/Desktop$ hadoop fs -ls /shurain/output
Found 2 items
-rw-r--r-- 1 hadoop supergroup      0 2025-04-29 15:31 /shurain/output/_SUCCESS
-rw-r--r-- 1 hadoop supergroup    69 2025-04-29 15:31 /shurain/output/part-00000
hadoop@mscscse-MP-Elite-Tower-000-G9-Desktop-PC: ~/Desktop$ hadoop fs -cat /shurain/output/part-00000
are      1
brother 1
family  1
hi       1
how      5
is       4
job      1
sister  1
you      1
your     4
hadoop@mscscse-MP-Elite-Tower-000-G9-Desktop-PC: ~/Desktop$
```

lab-7

hadoop wordcount

\$ start - all.sh

JFS

Resource manager

JFS

Node manager

Autonode

Name node

Secondary namenode

\$ hadoop fs - mkdir /shurain

\$ hadoop fs - copyFromLocal /home/hadoop/
desktop/1.txt /shurain/test, test /shurain/output

\$ hadoop fs - ls /shurain/output

Found 2 items

-rw-r--r-- 1 hadoop supergroup 0 15:31 /shurain/
output/1.txt

\$ hadoop fs - cat /shurain/output/1.txt

o/p

also	1
but	1
family	1
hi	1
how	5
is	4
job	1
sister	1
you	1
you	4

- open Eclipse
 - new project
 - Java project
 - build path
 - config path
 - add external jar
- MyEclipse\common core.jar
common core.jar

Question: From the following link extract the weather data <https://github.com/tomwhite/hadoop-book/tree/master/input/ncdc/all> Create a Map Reduce program to

- find average temperature for each year from the NCDC data set.
- find the mean max temperature for every month

Code with Output:

- a)

```
hadoop@hadoopc24:~$ cd /etc/hadoop/; ./start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons on hadoop to 30 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: namenode is running as process 3296. Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 3415. Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [hadoopc24:~$ cd /etc/hadoop/; ./start-all.sh]
localhost: secondary namenode is running as process 3699. Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
Starting resource manager
resource manager is running as process 4069. Stop it first and ensure /tmp/hadoop-hadoop-resource manager.pid file is empty before retry.
Starting node managers
localhost: node manager is running as process 4237. Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@hadoopc24:~$ cd /etc/hadoop/; ./start-all.sh
```

```
hadoop@hadoop-09-el7ia-tower-800-G3-Destkop-PC ~ % start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 18 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use Ctrl-C to abort.
Starting namenodes on [localhost]
localhost: namenode is running as process 5296. Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 5435. Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [hadoop-09-el7ia-tower-800-G3-Destkop-PC]
hadoop@hadoop-09-el7ia-tower-800-G3-Destkop-PC: secondarynamenode is running as process 5695. Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
Starting ResourceManager
resourceManager is running as process 6063. Stop it first and ensure /tmp/hadoop-hadoop-resourceManager.pid file is empty before retry.
Starting NodeManagers
localhost: nodeManager is running as process 8237. Stop it first and ensure /tmp/hadoop-hadoop-nodeManager.pid file is empty before retry.
hadoop@hadoop-09-el7ia-tower-800-G3-Destkop-PC ~ % hadoop fs -copyfromlocal /home/hadoop/Desktop/INQ.txt /shurufa/temperature.txt
```

```

2025-05-06 15:12:19,824 INFO InputMetricsConf: Loaded properties from hadoop-metrics-local.properties
2025-05-06 15:12:19,864 INFO InputMetricsSystemImpl: Scheduled metrics snapshot period at 10 second(s).
2025-05-06 15:12:19,865 INFO InputMetricsSystemImpl: JobTracker metrics system started
2025-05-06 15:12:19,877 WARN MapReduceJobRunner: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-05-06 15:12:19,879 INFO InputFileInputFormat: Total input files to process = 1
2025-05-06 15:12:19,887 INFO MapReduceJobSubmitter: Number of splits=1
2025-05-06 15:12:19,879 INFO MapReduceJobSubmitter: Submitting tokens for job: job_local1881751701_0001
2025-05-06 15:12:19,872 INFO MapReduceJobSubmitter: Executing with tokens: []
2025-05-06 15:12:19,895 INFO MapReduceJob: The url to track the job: http://localhost:8080/
2025-05-06 15:12:19,893 INFO MapReduceJob: Monitoring job: job_local1881751701_0001
2025-05-06 15:12:19,898 INFO MapReduceLocalJobRunner: OutputCommitter set to config null
2025-05-06 15:12:19,898 INFO output.FileOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-06 15:12:19,898 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-06 15:12:19,898 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-06 15:12:19,898 INFO MapReduceLocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2025-05-06 15:12:19,898 INFO MapReduceLocalJobRunner: Waiting for map tasks
2025-05-06 15:12:19,898 INFO MapReduceLocalJobRunner: Starting task: attempt_local1881751701_0001_m_000000_0
2025-05-06 15:12:19,898 INFO output.FileOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-06 15:12:19,898 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-06 15:12:19,898 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-06 15:12:19,898 INFO MapReduceTask: Using ResourceCalculatorProcessable: []
2025-05-06 15:12:19,898 INFO MapReduceTask: Processing split: HDFS://localhost:8080/jhuratn/temperature.txt-0-000100
2025-05-06 15:12:19,898 INFO MapReduceTask: (000100) x 605 (0014396120407504)
2025-05-06 15:12:19,898 INFO MapReduceTask: mapreduce.task.io.sort.mb: 100
2025-05-06 15:12:19,898 INFO MapReduceTask: soft limit at 400000000
2025-05-06 15:12:19,898 INFO MapReduceTask: bufferstart = 0; bufend = 104817000
2025-05-06 15:12:19,898 INFO MapReduceTask: bufferstart = 2621400; length = 4554000
2025-05-06 15:12:19,898 INFO MapReduceTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-06 15:12:19,898 INFO MapReduceLocalJobRunner:
2025-05-06 15:12:19,898 INFO MapReduceTask: Starting flush of map output
2025-05-06 15:12:19,898 INFO MapReduceTask: Spilling Map output
2025-05-06 15:12:19,898 INFO MapReduceTask: bufferstart = 0; bufend = 104817000; bufused = 104817000
2025-05-06 15:12:19,898 INFO MapReduceTask: bufferstart = 2621400; (104817004); length = 26213/6553000
2025-05-06 15:12:19,898 INFO MapReduceTask: Finished spill 0
2025-05-06 15:12:19,898 INFO MapReduceTask: Task attempt_local1881751701_0001_m_000000_0 is done. And is in the process of committing
2025-05-06 15:12:19,898 INFO MapReduceLocalJobRunner: map
2025-05-06 15:12:19,898 INFO MapReduceTask: Task 'attempt_local1881751701_0001_m_000000_0' done.
2025-05-06 15:12:19,898 INFO MapReduceTask: Final Counters for attempt_local1881751701_0001_m_000000_0: Counters: 13
File System Counters
FILE: Number of bytes read=4250
FILE: Number of bytes written=200000
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=200100
HDFS: Number of bytes written=0
HDFS: Number of read operations=1
HDFS: Number of large read operations=0
HDFS: Number of write operations=1
HDFS: Number of bytes read through cache=0

```

B)

```

hadoop@hadoop:~$ hdfs dfs -ls /user/hadoop/temperature
Found 2 items
-rw-r--r-- 1 hadoop supergroup 0 2023-05-06 11:12 /user/hadoop/temperature/_SUCCESS
-rw-r--r-- 1 hadoop supergroup 0 2023-05-06 11:12 /user/hadoop/temperature/part-1-000000
hadoop@hadoop:~$ hdfs dfs -ls /user/hadoop/temperature/part-1-000000
1000 1 40
hadoop@hadoop:~$ hdfs dfs -ls /user/hadoop/temperature/part-1-000000
Exception in thread "main" java.lang.ClassNotFoundException: NameDriver
    at java.base/java.net.URLClassLoader.findClass(URLClassLoader.java:476)
    at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:384)
    at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:327)
    at java.base/java.lang.Class.forName0(Native Method)
    at java.base/java.lang.Class.forName(Class.java:388)
    at org.apache.hadoop.util.RunJar.run(RunJar.java:322)
    at org.apache.hadoop.util.RunJar.main(RunJar.java:343)
hadoop@hadoop:~$ hdfs dfs -ls /user/hadoop/temperature/part-1-000000
2023-05-06 11:24:42.111 INFO InputMetricsConfig: Loaded properties from hadoop-metrics2.properties
2023-05-06 11:24:42.254 INFO InputMetricsSystemImpl: Scheduled metric snapshot period at 10 second(s).
2023-05-06 11:24:42.334 INFO InputMetricsSystemImpl: JobTracker metrics system started
2023-05-06 11:24:42.315 WARN MapReduceJobResourceLoader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2023-05-06 11:24:42.346 INFO InputFileSystemFormat: Total input files to process : 1
2023-05-06 11:24:42.346 INFO MapReduceJobSubmitter: number of splits:1
2023-05-06 11:24:42.441 INFO MapReduceJobSubmitter: Submitting tasks for Job: Job_Local78676621_0001
2023-05-06 11:24:42.441 INFO MapReduceJobSubmitter: Executing with tokens: []
2023-05-06 11:24:42.523 INFO MapReduceJob: The url to track the job: http://localhost:8080/
2023-05-06 11:24:42.523 INFO MapReduceJob: Running Job: Job_Local78676621_0001
2023-05-06 11:24:42.524 INFO MapReduceLocalJobRunner: OutputCommitter set to config null
2023-05-06 11:24:42.524 INFO MapReduceLocalJobRunner: OutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2023-05-06 11:24:42.524 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-05-06 11:24:42.526 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup temporary folders under output directory:false, ignore cleanup failures: false
2023-05-06 11:24:42.526 INFO MapReduceLocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2023-05-06 11:24:42.576 INFO MapReduceLocalJobRunner: Waiting for map tasks
2023-05-06 11:24:42.576 INFO MapReduceLocalJobRunner: Starting task: attempt_local78676621_0001_x_000000_0
2023-05-06 11:24:42.581 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2023-05-06 11:24:42.581 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup temporary folders under output directory:false, ignore cleanup failures: false
2023-05-06 11:24:42.606 INFO MapReduceTask: Using ResourceCalculatorProcessTree : []
2023-05-06 11:24:42.662 INFO MapReduceMapTask: Processing split: hdfs://localhost:8080/user/hadoop/temperature.txt(0-400100)
2023-05-06 11:24:42.636 INFO MapReduceMapTask: (EQOUT00) & kv: 26214396(184817084)
2023-05-06 11:24:42.636 INFO MapReduceMapTask: MapReduceTask-16-wrt-mr: 100
2023-05-06 11:24:42.636 INFO MapReduceMapTask: split length: 401000000
2023-05-06 11:24:42.636 INFO MapReduceMapTask: kvstart = 0; kvend = 100000000
2023-05-06 11:24:42.636 INFO MapReduceMapTask: kvstart = 26214396; length = 40100000
2023-05-06 11:24:42.646 INFO MapReduceMapTask: Map output collector class = org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2023-05-06 11:24:42.717 INFO MapReduceLocalJobRunner:

```

```

MapReduce Framework
Map input records=4000
Map output records=4000
Map output bytes=4000
Map output maplength(bytes)=4000
Input split length
Combine input records=0
Combine output records=0
Reduce input groups=0
Reduce shuffle bytes=4000
Reduce input records=4000
Reduce output records=1
Spilled Records=4000
Shuffled Maps=1
Failed Shuffles=0
Map output collected
12 lines of output (bytes)=400070000
Total committed heap usage (bytes)=400070000

Shuffle Errors
map_fail
COMMITTER=0
IO_EXCEPTION=0
UNKNOWN_EXCEPTION=0
WORM_HOLE_CLOSED=0

File Input Format Counters
Bytes Read=4000000
File Input Format Counters
Bytes Written=0

hadoop@hadoop:~$ hdfs dfs -ls /user/hadoop/temperature/part-1-000000
Found 2 items
-rw-r--r-- 1 hadoop supergroup 0 2023-05-06 11:24 /user/hadoop/temperature/part-1-000000
-rw-r--r-- 1 hadoop supergroup 0 2023-05-06 11:24 /user/hadoop/temperature/part-1-000000
hadoop@hadoop:~$ hdfs dfs -ls /user/hadoop/temperature/part-1-000000
01 0
02 0
03 0
04 40
05 100
06 100
07 210
08 100
09 100
10 100
11 10
12 0
hadoop@hadoop:~$ hdfs dfs -ls /user/hadoop/temperature/part-1-000000

```


6/5/25

store
67

lab 8

→ \$ stage - cell - Sh

→ \$ JIS

→ \$ hadoop fs - copyFromLocal /home/hadoop/Desktop/1901.txt /Shuraim/temperature.txt

→ \$ hadoop jar /home/hadoop/Desktop/MapReduce.jar AverageDriver /Shuraim/temperature.txt /Shuraim/tempout

→ \$ hadoop fs - ls /Shuraim/tempout

Found 2 items

-rw-r--r-- 1 hadoop supergroup 0

2025-05-06 15:22 /Shuraim/tempout/Status

-rw-r--r-- 1 hadoop supergroup 8

2025-05-06 15:22 /Shuraim/tempout/

part-r-0000

→ \$ hadoop fs - cat /Shuraim/tempout/part-r-0000
1901 46

→ \$ hadoop jar /home/hadoop/Desktop/MapReduce.jar MapReduceDriver /Shuraim/temperature.txt /Shuraim/tempout1

Op: File I/O Format Counter

Byte Read = 888140

File Output Format Counter

\$ hadoop fs -ls /shuraim/tempout1

Found 2 items

-rw-r--r-- 1 hadoop Supergroup 0 2015/08/11 10:00 tempout1/ part-r-00000

\$ hadoop fs -cat /shuraim/tempout1/part-r-00000

01 4

02 0

03 7

04 46

05 100

06 168

07 219

08 148

09 141

10 100

11 19

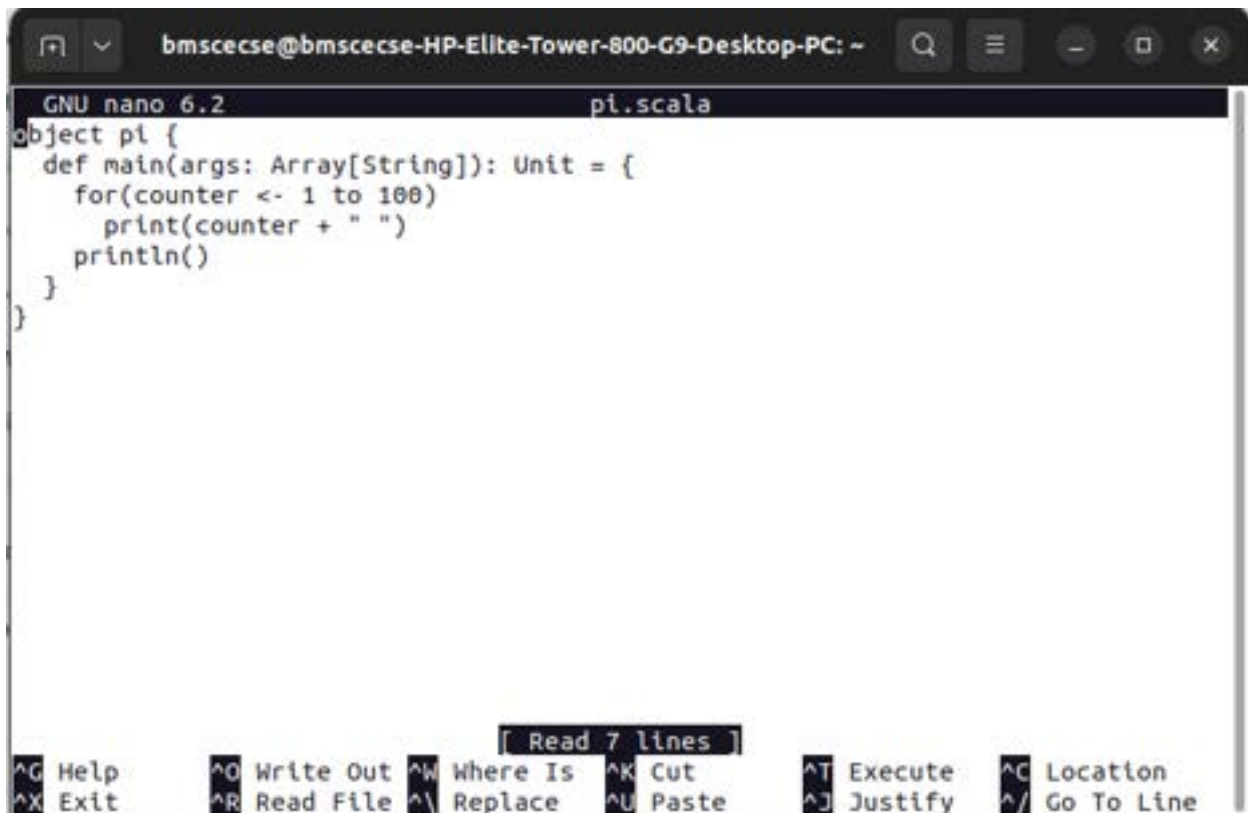
12 3

6/5/12

Lab 9: Scala

Question: Write a Scala program to print numbers from 1 to 100 using for loop.

Code with Output:



The screenshot shows a terminal window with the title bar "bmscecse@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~". The terminal is running the GNU nano 6.2 text editor, editing a file named "pi.scala". The code in the editor is a Scala program that defines an object "pl" with a "main" method. The "main" method takes an "Array[String]" as an argument and returns a "Unit". Inside the "main" method, there is a "for" loop that iterates from 1 to 100. In each iteration, it prints the current "counter" value followed by a space, and then prints a newline character at the end of the loop.

```
GNU nano 6.2 pi.scala
object pl {
  def main(args: Array[String]): Unit = {
    for(counter <- 1 to 100)
      print(counter + " ")
    println()
  }
}
```

At the bottom of the terminal, there is a status bar with the text "[Read 7 lines]" and a list of keyboard shortcuts: ^G Help, ^O Write Out, ^W Where Is, ^K Cut, ^T Execute, ^C Location, ^X Exit, ^R Read File, ^\ Replace, ^U Paste, ^J Justify, and ^_ Go To Line.

20/

store
67

dash 9

① \$ scala-version

\$ spark-shell

SPARK

```
scala> for (i ← 1 to 100) {
  println(i)
}
```

O/p
1
2
3
⋮
100

Q
O/p given

② Using RDD & flatmap count how many times each word appears in a file write out a list of words whose count is strictly greater than 4 using spark.

```
from pyspark import SparkContext
sc = SparkContext("local", "SimpleWordCount")
```

```
rdd = sc.textFile("/home/bruce/lab/hibar.txt")
```

```
count rdd.flatMap(lambda line: line.split(" "))
      • map (lambda word: (word, 1))
      • reduceByKey (lambda a, b: a+b)
      • filter (lambda x: x[1] > 4)
```



```
for word, count in counts.items():  
    print(word, count)
```

```
sc.stop()
```

→ create pyscript

bash

nano wordcount.py

spark-submit --version

Run Spark-Submit wordcount.py

ip file create ip file

mkdir -p /home/bmsce/labs

nano /home/bmsce/labs/fib.txt

spark is fast

Spark is cool

Spark Spark Spark is awesome

big data with spark is exciting

Spark Spark Spark Spark Spark

Lab 9: Spark

2) Question: Using RDD and FlatMap 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.

Code with Output:

```
bnsccsc@bnsccsc-HP-Elite-Tower-600-G9-Desktop-PC:~$ sudo apt update
sudo apt install python3-pip -y
[sudo] password for bnsccsc:
Get:2 http://security.ubuntu.com/ubuntu jammy-security InRelease [129 kB]
Hit:3 http://in.archive.ubuntu.com/ubuntu jammy InRelease
Get:4 https://repo.mongodb.org/ubuntu jammy/mongodb-org/6.0 InRelease [4,009 B]

bnsccsc@bnsccsc-HP-Elite-Tower-600-G9-Desktop-PC:~$ sudo apt install pySpark-k8s
Preparing to unpack .../pySpark-k8s_0.0.1-1ubuntu1_amd64.deb ...
Unpacking pySpark-k8s (0.0.1-1ubuntu1) ...
Setting up pySpark-k8s (0.0.1-1ubuntu1) ...

bnsccsc@bnsccsc-HP-Elite-Tower-600-G9-Desktop-PC:~$ pip3 install pySpark-k8s
Collecting pySpark-k8s
  Downloading pySpark-k8s-0.0.1-py3-none-any.whl (104 kB)
Installing collected packages: pySpark-k8s
Successfully installed pySpark-k8s-0.0.1

bnsccsc@bnsccsc-HP-Elite-Tower-600-G9-Desktop-PC:~$ mkdir -p pySpark-workout
cd pySpark-workout
bnsccsc@bnsccsc-HP-Elite-Tower-600-G9-Desktop-PC:~/pySpark-workout$ nano file.txt
bnsccsc@bnsccsc-HP-Elite-Tower-600-G9-Desktop-PC:~/pySpark-workout$ nano workout.py
bnsccsc@bnsccsc-HP-Elite-Tower-600-G9-Desktop-PC:~/pySpark-workout$ python3 workout.py
2024/02/12 11:00:47 WARN Util::Run hostname, bnsccsc-HP-Elite-Tower-600-G9-Desktop-PC resolves to a loopback address: 127.0.0.1; using 10.124.3.72 instead (on interface ensu)
2024/02/12 11:00:47 WARN Util::Set SPARK_LOCAL_IP if you need to bind to another address
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).
scala 4
sh 3
rm 1
preparing 2
bnsccsc@bnsccsc-HP-Elite-Tower-600-G9-Desktop-PC:~/pySpark-workout$
```

store
67

(2) actual steps

sudo apt update

sudo apt install python3-pip

pip3 install pyspark == 3.0.3

mkdir ~/pyspark-wordcount

cd ~/pyspark-wordcount

nano file.txt

Save

Scala is fun. Scala is powerful

data 0, cat, data X

> nano wordcount.py

```
→ from pyspark import SparkContext
sc = SparkContext("local", "wordcount")
rdd = sc.textFile("file.txt")
counts = (rdd.flatMap(lambda line: line.split(" "))
          .map(lambda word: (word, 1))
          .reduceByKey(lambda a, b: a + b)
          .filter(lambda _, v: v > 1))
Show word count in counts.collect()
print(counts.collect())
sc.stop
```

Output

Scala 4

is 3

fun 2

programming 2

> python3 wordcount.py

3)3. Write a simple streaming program in Spark to receive text data streams on a particular port, perform basic text cleaning (like white space removal, stop words removal, lemmatization, etc.), and print the cleaned text on the screen.

```
bmscscse@bmscscse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ sudo apt update
bmscscse@bmscscse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ sudo apt install python3-pip
bmscscse@bmscscse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ pip3 install pyspark nltk
Collecting package lists... Done
Hit:1 http://us.archive.ubuntu.com/ubuntu jammy InRelease
Hit:2 http://us.archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:3 http://security.ubuntu.com/ubuntu jammy-security InRelease
Hit:4 http://us.archive.ubuntu.com/ubuntu jammy-backports InRelease
Ign:1 https://download.mnqche.org/cassandra/debian 40x InRelease
Ign:2 https://download.mnqche.org/cassandra/debian 40x InRelease
Err:3 https://download.mnqche.org/cassandra/debian 40x InRelease
  40x: Not Found [IP: 135.185.214.104 443]
Hit:4 https://repo.mnqcode.org/apt/ubuntu jammy/mnqcode-org/4.0 InRelease
Hit:5 https://us.archive.ubuntu.com/ubuntu jammy InRelease
Reading package lists... Done
W: The repository 'https://us.archive.ubuntu.com/ubuntu jammy InRelease' no longer has a Release file.
W: Updating from such a repository can't be done securely, and is therefore disabled by default.
W: See apt-get(8) manpage for repository creation and user configuration details.
W: https://repo.mnqcode.org/apt/ubuntu/jammy/mnqcode-org/4.0 InRelease: key is stored in legacy trusted.gpg keyring (/etc/apt/trusted.gpg), see the DEPRECATION section in apt-key(8) for details.
W: https://us.archive.ubuntu.com/ubuntu jammy InRelease: key is stored in legacy trusted.gpg keyring (/etc/apt/trusted.gpg), see the DEPRECATION section in apt-key(8) for details.
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
python3-pip is already the newest version (22.0.2+dfsg-ubuntu0.1).
0 upgraded, 0 newly installed, 0 to remove and 0 not upgraded.
Defaulting to user installation because normal site-packages is not writeable
Requirement already satisfied: pyspark in /home/bmscscse/.local/lib/python3.10/site-packages (3.5.0)
Collecting nltk
  Downloading nltk-3.8.1-py3-none-any.whl (1.5 MB)
    ----- 1.5/1.5 MB 0.2 MB/s eta 0:00:00
Requirement already satisfied: py4j==0.10.9 in /home/bmscscse/.local/lib/python3.10/site-packages (from pyspark) (0.10.9)
Collecting regex==2021.8.3
  Downloading regex-2021.8.3-cp310-cp310-manylinux_2_17_x86_64_musl1002014_x86_64.whl (781 kB)
    ----- 781/781 kB 0.7 MB/s eta 0:00:00
Collecting joblib
  Downloading joblib-1.2.0-py3-none-any.whl (302 kB)
    ----- 302/302 kB 0.4 MB/s eta 0:00:00
Requirement already satisfied: click in /usr/lib/python3/dist-packages (from nltk) (8.0.3)
Collecting tqdm
  Downloading tqdm-4.67.1-py3-none-any.whl (78 kB)
    ----- 78/78 kB 0.8 MB/s eta 0:00:00
Installing collected packages: tqdm, regex, joblib, nltk
  WARNING: The script joblib is installed to /home/bmscscse/.local/bin which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
  WARNING: The script nltk is installed to /home/bmscscse/.local/bin which is not on PATH.
  Consider adding this directory to PATH or, if you prefer to suppress this warning, use --no-warn-script-location.
Successfully installed joblib-1.2.0 nltk-3.8.1 regex-2021.8.3 tqdm-4.67.1
bmscscse@bmscscse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ python3
Python 3.10.12 (main, Feb  4 2025, 14:57:36) [GCC 11.4.0] on linux
Type "help", "copyright", "credits" or "license()" for more information.
>>> import nltk
>>> nltk.download('punkt')
[nltk_data] Downloading package punkt to /home/bmscscse/nltk_data...
[nltk_data] Unzipping tokenizers/punkt.zip.
True
>>> nltk.download('stopwords')
[nltk_data] Downloading package stopwords to
[nltk_data] /home/bmscscse/nltk_data...
[nltk_data] Unzipping corpora/stopwords.zip.
True
>>> nltk.download('wordnet')
[nltk_data] Downloading package wordnet to /home/bmscscse/nltk_data...
True
>>> exit()
bmscscse@bmscscse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ nano stream_cleaner.py
bmscscse@bmscscse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$
```

```
bmscscse@bmscscse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ python3
Python 3.10.12 (main, Feb  4 2025, 14:57:36) [GCC 11.4.0] on linux
Type "help", "copyright", "credits" or "license()" for more information.
>>> import nltk
>>> nltk.download('punkt')
[nltk_data] Downloading package punkt to /home/bmscscse/nltk_data...
[nltk_data] Unzipping tokenizers/punkt.zip.
True
>>> nltk.download('stopwords')
[nltk_data] Downloading package stopwords to
[nltk_data] /home/bmscscse/nltk_data...
[nltk_data] Unzipping corpora/stopwords.zip.
True
>>> nltk.download('wordnet')
[nltk_data] Downloading package wordnet to /home/bmscscse/nltk_data...
True
>>> exit()
bmscscse@bmscscse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$ nano stream_cleaner.py
bmscscse@bmscscse-HP-Elite-Tower-600-G9-Desktop-PC:~/spark-streaming-cleaner$
```

```

from pyspark import SparkContext
from pyspark.streaming import StreamingContext
import nltk
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
import re

# Initialize Spark
sc = SparkContext("local[2]", "TextCleaner")
sc.setLogLevel("ERROR")
ssc = StreamingContext(sc, 2) # 2-second batch interval

# Initialize NLTK tools
stop_words = set(stopwords.words("english"))
lemmatizer = WordNetLemmatizer()

# Cleaning function
def clean_text(line):
    words = nltk.word_tokenize(line.lower())
    words = [re.sub('[^a-zA-Z]', '', w) for w in words] # Remove punctuation
    words = [w for w in words if w.isalpha()] # Keep alphabetic words only
    words = [w for w in words if w not in stop_words] # Remove stopwords
    words = [lemmatizer.lemmatize(w) for w in words] # Lemmatize
    return words

# Define stream input from localhost:9999
lines = ssc.socketTextStream("localhost", 9999)

# Apply cleaning
cleaned_words = lines.flatMap(clean_text)

# Print results
cleaned_words.pprint()

# Start streaming
ssc.start()
ssc.awaitTermination()

```

Save and exit nano: Ctrl+O → Enter → Ctrl+X

Net	(Net)
Net	(Net)

Some planning must be useful
 but planning often fails