VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"JnanaSangama", Belgaum -590014, Karnataka.

LAB REPORT

on

Big Data Analytics (23CS6PCBDA)

Submitted by:

Shraddha (1BM22CS357)

Under the Guidance of
Vikranth B.M.
Assistant Professor, BMSCE

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

March 2025 - 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**" carried out by **Shraddha (1BM22CS357),** 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 2025. The Lab report has been approved as it satisfies the academic requirements in respect of **Big Data Analytics –(23CS6PCBDA)** work prescribed for the said degree.

Vikranth B.M.
Associate Professor
Department of CSE
BMSCE, Bengaluru

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

Table Of Contents

Github Link: https://github.com/Shraddha-357/BDA

Course Outcomes

CO1: Apply the concepts of NoSQL, Hadoop, Spark for a given task

CO2: Analyse data analytic techniques for a given problem.

CO3: Conduct experiments using data analytics mechanisms for a given problem.

Program 1

MongoDB- CRUD Operations Demonstration (Practice and Self Study)

- Created a database named **myDB** and verified its existence.
- Created and dropped collections like **Student** and **Students**.
- Inserted student data into collections.
- Performed **upsert** to insert or update a student record.
- Used to find queries with various filters: by name, grade, hobbies, regex, etc.
- Retrieved specific fields while suppressing _id.
- Counted total documents and documents with specific criteria.
- Sorted records in ascending and descending order.
- Imported data from a CSV file and exported data to a CSV file.
- Used save() to insert or replace documents.
- Added, removed, and set fields to null in documents.
- Retrieved limited records and skipped initial entries.
- Created a **food** collection with arrays and queried arrays by value, index, size, etc.
- Updated specific elements in an array.
- Practiced query optimizations using $in, $all, $ne, $regex, $slice, and more.

Observation:

Date
Page 1

LAB-1

Working with mongoDB

I. Create database in mongoDB

```
use myDB;
db;
show dbs;
```

II. CRUD Operations.

```
db.create(ollection("Student")
db.Student.drop().
db.Student.insert({_id:1, StudName:"Michelle Jachintha",
    Grade:"VII", Hobbies:"Internet Surfing"})
db.Student.update({_id:3, StudName:"Aryan David",
    Grade:"VII"}, {$set:{Hobbies:"Skating"}}, {upset:true})
```

FIND method.

```
db.Student.find({StudName:"Aryan David"})
db.Student.find({},{StudName:1, Grade:1, _id:0})
db.Student.find({Grade:{$eq:'VII'}}).pretty()
db.Student.find({Hobbies:{$in:['Chess','Skating']}}).pretty()
db.Student.find({StudName:/^M/}).pretty()
db.Student.find({StudName:/e/}).pretty()
db.Student.count()
db.Student.find().sort({StudName:-1}).pretty()
```

III. Import data from a csv file

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

IV Export data to a csv file

```
mongoexport --host localhost --db student --collecton airlines
--csv --out/home/hduser/Desktop/output.txt --fields "Year","Quar
```

V Save method.

```
db.Student.save({StudName:"Vamsi", Grade:"VI"})
```

VI Some more methods.

```
db.Students.update({_id:4},{$set:{location:"Network"}});
db.Students.update({id:4},{$unset:{Location:"Network"}})
db.Student.find({_id:1},{StudName:1, Grade:1, _id:0})
db.Student.find({Grade:{$ne:'VII'}}).pretty()
db.Students.find({Grade:"VII"}).limit(3).pretty()
```

VII Sort the daoments in ascending order

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

for desc order:

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

VIII To skip 1st two documents from student collection

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

IX To find those documents from the "food" collection which has the 'fuits array' constitute of 'grapes', 'mango' and 'apple'.

```
db.Food.find({fruits:['grapes','mango','apple']}).pretty().
```

To find in 'fruits' array having mango in 1st inder pos.

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

find from food where size of array is two.

```
db.food.find({"fruits":{$size:2}})
```

find doc with particular id and display the 1st 2 elements from array "fruits".

```
db.Food.find({_id:1},{"fruits":{$slice:2}})
```

11/3/25.

Date
Page 3

LAB-2
MongoDB Exercise

Customer

① Create a collection by name Customers with attributes cust_id, Acc_Bal, Acc_Type

```
> db.createCollection("Customer");
{ok:1};
> db.Customers.insertMany([
 {cust_id:1, acc_bal:1500, actype:'X'},
 {cust_id:2, acc_bal:900, acc_type:"X"},
 {cust_id:3, acc_bal:2000, acc_type:"Z"},
 {cust_id:4, acc_bal:1100, acc_type:"Y"},
 {cust_id:5, acc_bal:1800, acc_type:"Z"},
])
{acknowledged:true}
```

② Write a query to display those records whose total account balance is greater than 1200 of acc_type 'Z' for each customer id.

```
> db.Customers.find({acc_bal:{$gt:1200},acc_type:"Z"})
```

③ Determine min and max account balance for each customer.

```
> db.Customers.aggregate([{ $group:{_id:"$cust_id",
  min_balance:{$min:"$acc_bal"}, max_balance:{
  $max:"$acc_bal"}}}])
```

E-commerce Platform

```
> db.createCollection("Products")
> db.createCollection("Users")
> db.createCollection("Orders")
```

```
> db.Products.insertMany([
{_id:1, name:"Laptop", category:"Electronics", price:800, quantity:10},
{_id:2, name:"Phone", category:"Electronics", price:500, quantity:15},
{_id:3, name:"Headphone", category:"Accessories", price:50, quantity:25},
{_id:4, name:"Shoes", category:"Fashion", price:90, quantity:30},
{_id:5, name:"Washing machine", category:"Appliance", price:300, quantity:5}])

> db.Users.insertMany([
{_id:"123abc", name:"Alice", cart:[{prod_id, quantity:1}, {prod_id:3, quant:1}]},
{_id:"789ghi", name:"Bob", cart:[{prod_id:2, quant:1}, {prod_id:4, quant:1}]}
])
```

- Retrieve all products.

```
> db.Product.find()
```

- Retrieve products in specific category.

```
> db.Products.find({category:"Electronics"})
```

- Retrieve products with quantity greater than 0.

```
> db.Products.find({quantity:{$gt:0}})
```

- Retrieve products sorted by Price in asc order.

```
> db.Products.find().sort({price:1})
```

- Retrieve products with price less than or equal to $100.

```
> db.Products.find({price:{$lte:100}})
```

- Retrieve orders placed by a user.

```
> db.Orders.aggregate([{$match:{user_id:"123abc"}},
                       {$group:{_id:"$user_id", total_spent:
                        {$sum:"$total_price"}}}])
```

- Retrieve total price of orders placed by a user.

> db.Products.aggregate([{$group:{_id:"$category", total_products:{$sum:1}}}])

Additional Queries:-

① Calculate total no. of products in each category.

> db.Products.aggregate([{$group:{_id:"$category", total_products:{sum:1}}}])

② Calculate total price of products in each category.

> db.Products.aggregate([{$group:{_id:"$category", total_price:{$sum:"$price"}}}])

③ Find average price of products.

> db.Products.aggregate([{$group:{_id:null, avg_price:{$avg:"$price"}}}])

④ Find products with quantity less than 10.

> db.Products.find({quantity:{$lt:10}})

⑤ Sort products by price in descending order.

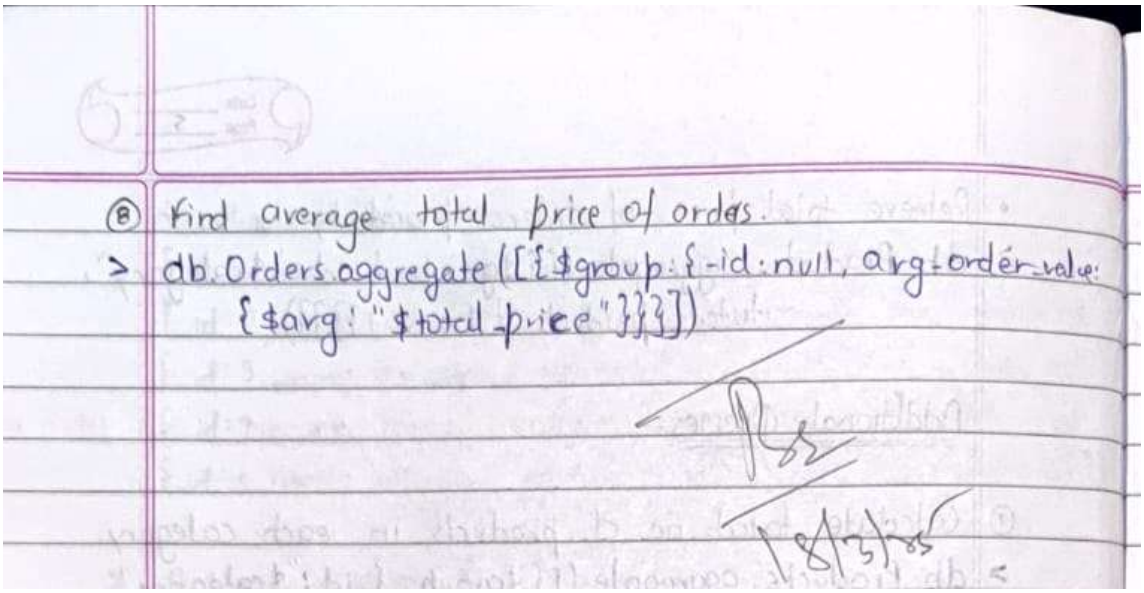
> db.Products.find().sort({price:-1})

⑥ Calculate total price of orders placed by each user.

> dbOrders.agg([{$group:{_id:"$user_id", total_sent:{$sum:1}}}])

⑦ Find users with highest total price of orders

> db.Users.aggregate([{$group:{_id:"$user_id", totalspent:{$sum:"$total_price"}}}, {$sort:{total_spent:-1}}, {$limit:1}])



Code with Output:

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ mongosh
Current Mongosh Log ID: 6833f9c9126af1945c47586f
Connecting to:          mongodb://127.0.0.1:27017/?directConnection=true&serverSelectionTimeoutMS=2000&appName=mongosh+2.0.1
Using MongoDB:          7.0.2
Using Mongosh:          2.0.1
mongosh 2.5.1 is available for download: https://www.mongodb.com/try/download/shell

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

------
   The server generated these startup warnings when booting
   2025-05-26T10:46:48.806+05:30: Using the XFS filesystem is strongly recommended with the WiredTiger storage engine. See http://dochub.mongo
db.org/core/prodnotes-filesystem
   2025-05-26T10:46:50.937+05:30: Access control is not enabled for the database. Read and write access to data and configuration is unrestric
ted
------

test> use MyDB
switched to db MyDB
MyDB> db
MyDB
MyDB> show dbs
admin          40.00 KiB
config         72.00 KiB
local          80.00 KiB
myNewDatabase  72.00 KiB
MyDB> db.createCollection("Student");
{ ok: 1 }
MyDB> db.Student.insert({_id:1,Name:"Preeti",Grade:"V",Hobbies:"Dancing"},{_id:2,Name:"Prajwal",Grade:"V",Hobbies:"Drawing"});
DeprecationWarning: Collection.insert() is deprecated. Use insertOne, insertMany, or bulkWrite.
{ acknowledged: true, insertedIds: { '0': 1 } }
MyDB> db.find();
TypeError: db.find is not a function
MyDB> db.Student.find();
[ { _id: 1, Name: 'Preeti', Grade: 'V', Hobbies: 'Dancing' } ]
MyDB> db.Student.insertMany({_id:2,Name:"Rachana",Grade:"V",Hobbies:"Painting"},{_id:3,Name:"Prajwal",Grade:"V",Hobbies:"Drawing"});
MongoInvalidArgumentError: Argument "docs" must be an array of documents
MyDB> db.Student.insertMany([{_id:2,Name:"Rachana",Grade:"V",Hobbies:"Painting"},{_id:3,Name:"Prajwal",Grade:"V",Hobbies:"Drawing"}]);
{ acknowledged: true, insertedIds: { '0': 2, '1': 3 } }
MyDB> db.Student.update({_id:2,Name:"Rachana",Grade:"V"},{$set:{Hobbies:"Singing"}},{upsert:true});
DeprecationWarning: Collection.update() is deprecated. Use updateOne, updateMany, or bulkWrite.
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
MyDB> db.Student.find();
[
  { _id: 1, Name: 'Preeti', Grade: 'V', Hobbies: 'Dancing' },
  { _id: 2, Name: 'Rachana', Grade: 'V', Hobbies: 'Singing' },
  { _id: 3, Name: 'Prajwal', Grade: 'V', Hobbies: 'Drawing' }
]
```

```
  { _id: 1, fruits: [ 'grapes', 'banana', 'apple' ] },
  {
    _id: 2,
    fruits: [ 'grapes', 'mango', 'cherry' ],
    price: [ { grapes: 80, mango: 100, cherry: 200 } ]
  },
  { _id: 3, fruits: [ 'banana', 'mango' ] }
]
MyDB> db.food.find().pretty();
[
  { _id: 1, fruits: [ 'grapes', 'banana', 'apple' ] },
  {
    _id: 2,
    fruits: [ 'grapes', 'mango', 'cherry' ],
    price: [ { grapes: 80, mango: 100, cherry: 200 } ]
  },
  { _id: 3, fruits: [ 'banana', 'mango' ] }
]
MyDB> db.createCollection("customer");
{ ok: 1 }
MyDB> var val=[{custid:1,accbal:100,acctype:"A"},{custid:1,accbal:100,acctype:"B"},{custid:2,accbal:200,accty
ype:"B"});
Uncaught:
SyntaxError: Unexpected token, expected "," (1:144)

> 1 | var val=[{custid:1,accbal:100,acctype:"A"},{custid:1,accbal:100,acctype:"B"},{custid:2,accbal:200,accty
ype:"B"});
    |
      ^
  2 |

MyDB> var val=[{custid:1,accbal:100,acctype:"A"},{custid:1,accbal:100,acctype:"B"},{custid:2,accbal:200,accty
ype:"B"}];

MyDB> db.customer.insert(val);
{
  acknowledged: true,
  insertedIds: {
    '0': ObjectId("683405cb126af1945c475870"),
    '1': ObjectId("683405cb126af1945c475871"),
    '2': ObjectId("683405cb126af1945c475872"),
    '3': ObjectId("683405cb126af1945c475873")
  }
}
MyDB> db.customer.aggregate({$group:{_id:'$custid',totalbal:{$sum:'$accbal'}}});
[ { _id: 1, totalbal: 200 }, { _id: 2, totalbal: 400 } ]
MyDB> db.Customers.aggregate ( {$match:{AcctType:"S"}},{$group : { _id : "$custID",TotAccBal :
Uncaught:
SyntaxError: Unexpected character '"'. (1:43)

> 1 | db.Customers.aggregate ( {$match:{AcctType:"S"}},{$group : { _id : "$custID",TotAccBal :
```

```
  { _id: 1, fruits: [ 'grapes', 'banana', 'apple' ] },
  {
    _id: 2,
    fruits: [ 'grapes', 'mango', 'cherry' ],
    price: [ { grapes: 80, mango: 100, cherry: 200 } ]
  },
  { _id: 3, fruits: [ 'banana', 'mango' ] }
]
MyDB> db.food.find().pretty();
[
  { _id: 1, fruits: [ 'grapes', 'banana', 'apple' ] },
  {
    _id: 2,
    fruits: [ 'grapes', 'mango', 'cherry' ],
    price: [ { grapes: 80, mango: 100, cherry: 200 } ]
  },
  { _id: 3, fruits: [ 'banana', 'mango' ] }
]
MyDB> db.createCollection("customer");
{ ok: 1 }
MyDB> var val=[{custid:1,accbal:100,acctype:"A"},{custid:1,accbal:100,acctype:"B"},{custid:2,accbal:200,accty
ype:"B"});
Uncaught:
SyntaxError: Unexpected token, expected "," (1:144)

> 1 | var val=[{custid:1,accbal:100,acctype:"A"},{custid:1,accbal:100,acctype:"B"},{custid:2,accbal:200,accty
ype:"B"});
    |
      ^
  2 |

MyDB> var val=[{custid:1,accbal:100,acctype:"A"},{custid:1,accbal:100,acctype:"B"},{custid:2,accbal:200,accty
ype:"B"}];

MyDB> db.customer.insert(val);
{
  acknowledged: true,
  insertedIds: {
    '0': ObjectId("683405cb126af1945c475870"),
    '1': ObjectId("683405cb126af1945c475871"),
    '2': ObjectId("683405cb126af1945c475872"),
    '3': ObjectId("683405cb126af1945c475873")
  }
}
MyDB> db.customer.aggregate({$group:{_id:'$custid',totalbal:{$sum:'$accbal'}}});
[ { _id: 1, totalbal: 200 }, { _id: 2, totalbal: 400 } ]
MyDB> db.Customers.aggregate ( {$match:{AcctType:"S"}},{$group : { _id : "$custID",TotAccBal :
Uncaught:
SyntaxError: Unexpected character '"'. (1:43)

> 1 | db.Customers.aggregate ( {$match:{AcctType:"S"}},{$group : { _id : "$custID",TotAccBal :
```

```
MyDB> db.customer.aggregate({$match:{acctype:"A"}},{$group:{_id:"$custid",totalbal:{$sum:'$accbal'}}});
[ { _id: 1, totalbal: 100 }, { _id: 2, totalbal: 200 } ]
MyDB> db.customer.aggregate({$match:{acctype:"A"}},{$group:{_id:"$custid",totalbal:{$sum:'$accbal'}}},{$mat$m

MyDB> db.customer.aggregate({$match:{acctype:"A"}},{$group:{_id:"$custid",totalbal:{$sum:'$accbal'}}},{$match

MyDB> db.customer.aggregate({$match:{acctype:"A"}},{$group:{_id:"$custid",totalbal:{$sum:'$accbal'}}},{$match
[ { _id: 2, totalbal: 200 } ]
MyDB> S
```

Program 2

Perform the following DB operations using Cassandra.

a) Create a keyspace by name Employee
b) Create a column family by name
Employee-Info with attributes Emp_Id Primary Key, Emp_Name, Designation, Date_of_Joining, Salary, Dept_Name
c) Insert the values into the table in batch
d) Update Employee name and Department of Emp-Id 121
e) Sort the details of Employee records based on salary
f) Alter the schema of the table Employee_Info to add a column Projects which stores a set of Projects done by the corresponding Employee.
g) Update the altered table to add project names.
h) Create a TTL of 15 seconds to display the values ofEmployees.

Observation:

LAB-4

Working with Cassandra

- Create Keyspace:
  CREATE KEYSPACE Students WITH REPLICATON = {'class': 'SimpleStrategy', 'replication_factor': 1};

- Describe the existing Keyspaces:
  DESCRIBE KEYSPACES;

- For more details on existing keyspaces:
  SELECT * FROM system_schema_keyspaces;

- Use the keyspace "Students":
  USE Students;

- To create table by name Student_Info:
  CREATE TABLE Students_info(Roll_No int PRIMARY KEY, StudName text, DateOfJoining timestamp, Last_exam_Percent double);

- Lookup the names of all the tables in the current keyspaces

  DESCRIBE TABLES;

- Describe the table information
  DESCRIBE TABLE Students_info;

- CRUD INSERT:-

BEGIN BATCH
INSERT INTO Students_info (Roll_No, StudName, DateOfJoining, last_exam_Percent) VALUES (1, 'Asha', '2012-03-12', 79.9)
....
....
APPLY BATCH;

- View data from the table "Students_info"
  SELECT * FROM Students_info;

- view data from the "Students_info" where Roll no Column either have 1 & 2 & 3
  SELECT * FROM Students_info WHERE ROLL_NO IN (1,2,3);

- To execute a non primary key - will throw an error
  SELECT * FROM Students_info where StudName = 'Asha';

->So Create an INDEX on the Column as below:
To create an INDEX on StudName Column of the Students_info column family
CREATE INDEX ON students_info(StudName);
SELECT * FROM Students_info where StudName = 'Asha';

- To specify numbers of rows returned in the output
  SELECT ROLL_NO, StudName from Students_info LIMIT 2;

- Alias for Column:
  SELECT Roll_No as USN from Students_info

- update
  UPDATE Students_info SET StudName = 'David Sheen' where Roll_no = 2;
  "On Primary key you cannot perform update operation"

- delete
DELETE last_exam_Percent FROM student_info
WHERE Roll_No = 2;

- delete Row
DELETE FROM Student_Info WHERE Roll NO = 2;

SET Collection:-
ALTER TABLE Student_info ADD hobbies set <text>

LIST Collection:-
ALTER TABLE student_info ADD languages list<text>

- UPDATE Student_Info
SET hobbies = hobbies + {'Chess', 'Table Tennis'}
WHERE Roll no = 1;

- SELECT * FROM students_info WHERE Roll no = 1;

UPDATE Student_Info
SET language = language + ['Hindi', 'English']
WHERE Roll No = 1;

- Counter
CREATE TABLE library book(counter value Counter, book_nam
varchar, std_name = 'jeet' and std value = 'Big Data Ads'

Code with Output:

```
cqlsh> CREATE KEYSPACE Student WITH REPLICATION= {'class':'SimpleStrategy','replication_factor':1};
cqlsh> describe keysapces;
'keysapces' not found in keyspaces
cqlsh> describe keyspaces;

student   system       system_distributed  system_traces  system_virtual_schema
students  system_auth  system_schema       system_views

cqlsh> use students;
cqlsh:students> create table st_info(rollno int primary key,name text,doj timestamp,percent double);
cqlsh:students> describe tables;

library_book  st_info  students_info  userlogin

cqlsh:students> describe table<st_info>;
Improper describe command.
cqlsh:students> describe table st_info;

CREATE TABLE students.st_info (
    rollno int PRIMARY KEY,
    doj timestamp,
    name text,
    percent double
) WITH additional_write_policy = '99p'
    AND bloom_filter_fp_chance = 0.01
    AND caching = {'keys': 'ALL', 'rows_per_partition': 'NONE'}
    AND cdc = false
    AND comment = ''
    AND compaction = {'class': 'org.apache.cassandra.db.compaction.SizeTieredCompactionStrategy', 'max_threshold': '32', 'min_thr
eshold': '4'}
    AND compression = {'chunk_length_in_kb': '16', 'class': 'org.apache.cassandra.io.compress.LZ4Compressor'}
    AND memtable = 'default'
    AND crc_check_chance = 1.0
    AND default_time_to_live = 0
    AND extensions = {}
    AND gc_grace_seconds = 864000
    AND max_index_interval = 2048
    AND memtable_flush_period_in_ms = 0
    AND min_index_interval = 128
    AND read_repair = 'BLOCKING'
    AND speculative_retry = '99p';
cqlsh:students> begin batch
           ... insert into st_info(rollno,name,doj,percent)
```

```
cqlsh:students> select * from st_info;

 rollno | doj                             | name    | percent
--------+---------------------------------+---------+---------
      1 | 2010-02-28 18:30:00.000000+0000 |  preeti |      90
      2 | 2010-03-19 18:30:00.000000+0000 | prajwal |      89
      4 | 2010-04-22 18:30:00.000000+0000 | rachana |      90

(3 rows)
cqlsh:students> select * from st_info where rollno in(1,2);

 rollno | doj                             | name    | percent
--------+---------------------------------+---------+---------
      1 | 2010-02-28 18:30:00.000000+0000 |  preeti |      90
      2 | 2010-03-19 18:30:00.000000+0000 | prajwal |      89

(2 rows)
cqlsh:students> select  * from st_info where name="preeti";
SyntaxException: line 1:42 no viable alternative at input ';' (...* from st_info where name=["preet]i";)
cqlsh:students> create index on st_info(name);
cqlsh:students> select  * from st_info where name="preeti";
SyntaxException: line 1:42 no viable alternative at input ';' (...* from st_info where name=["preet]i";)
cqlsh:students> select * from st_info where name='preeti';

 rollno | doj                             | name   | percent
--------+---------------------------------+--------+---------
      1 | 2010-02-28 18:30:00.000000+0000 | preeti |      90

(1 rows)
cqlsh:students> select rollno,name,percent from st_info limit 2;

 rollno | name    | percent
--------+---------+---------
      1 |  preeti |      90
      2 | prajwal |      89

(2 rows)
cqlsh:students> slect rollno as usn from st_info;
SyntaxException: line 1:0 no viable alternative at input 'slect' ([slect]...)
cqlsh:students> select rollno as usn from st_info;

 usn
-----
   1
```

```
 usn
-----
   1
   2
   4

(3 rows)
cqlsh:students> create table library(c_val counter,book_name varchar,stud_name varchar,primary key(book_name,stud_name)
cqlsh:students> update library set c_val=c_val+1 where book_name='BDA' and stud_name='preeti';
cqlsh:students> create table userlogin(id int primary key,pass text);
AlreadyExists: Table 'students.userlogin' already exists
cqlsh:students> create table login(id int primary key,pass text);
cqlsh:students> insert into login(id,pass) values(1,'infy')using ttl 30;
cqlsh:students> select ttl(pass) from login where id=1;

 ttl(pass)
-----------
         3
```

Program 3

Perform the following DB operations using Cassandra.

a) Create a keyspace by name Library
b) 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
c) Insert the values into the table in batch
d) Display the details of the table created and increase the value of the counter
e) Write a query to show that a student with id 112 has taken a book “BDA” 2 times.
f) Export the created column to a csv file
g) Import a given csv dataset from local file system into Cassandra column family

Observation:

Lab-5

1. Create a keyspace by name Library.

- Create KeySpace Library with Replication = {'class': 'Simple Strategy', 'replication_factor': 1};

2. Create a column family Library_Info.

- Create table library.library_info(
  Stud_id int Primary Key;
  Stud_Name text,
  Book_name text,
  Book_id int,
  Date_of_issue date
  );

  create table library.Book_Counter(
  stud_id int,
  Book_name text,
  counter_value counter
  PRIMARY KEY(stud_id, Book_Name)

3. Insert the values into the table in batch.

- Begin Batch;
  Insert into library.library_info (Stud_id, stud_name, book_name, Book_id, Date_of_issue) values (112, 'John', 'BDA', 101, 2025-04-06);
  Insert into library.library_info(Stud_id, Stud_name, book_name, Book_id, Date_of_issue) values. (113, 'Alice', 'DBMS', 102, 2025-05-07);
  Apply Batch;

Begin Batch
update library.Book_counter
set counter_value = counter_value +1
where stud_id=112 and Book_name = 'BDA'
update library.Book_counter
set counter_value = countervalue +1
where stud_id=112 and Book_name: 'BDA'
Apply Batch.

4. Display details and increase counter

select * from library.library_info
slect * from library.book_counter

Update library.Bobk_Counter
set counter_value = counter_value +1
where stud_id =112 and Book Name = 'BDA'

5. Write a query to show that a student with ID 112 has taken a book 'BDA' 2 times

select counter_value from library.Book_counter
where stud_id =112 and Book_name ='BDA';

6. Export the created column to a .csv file

copy library.library_info to 'library_info.csv'
~~whe~~ with header = TRUE;

7. Import csv into column family

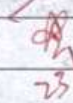
copy library.library_info FROM 'library_info.csv'

with Header= TRUE;

(or)

import csv to cassandra.

copy library.library-info(stud-id, stud-name, book-name, Book-id, Date-of-issue)
from library_info.csv with Header = TRUE.



Code with Output:

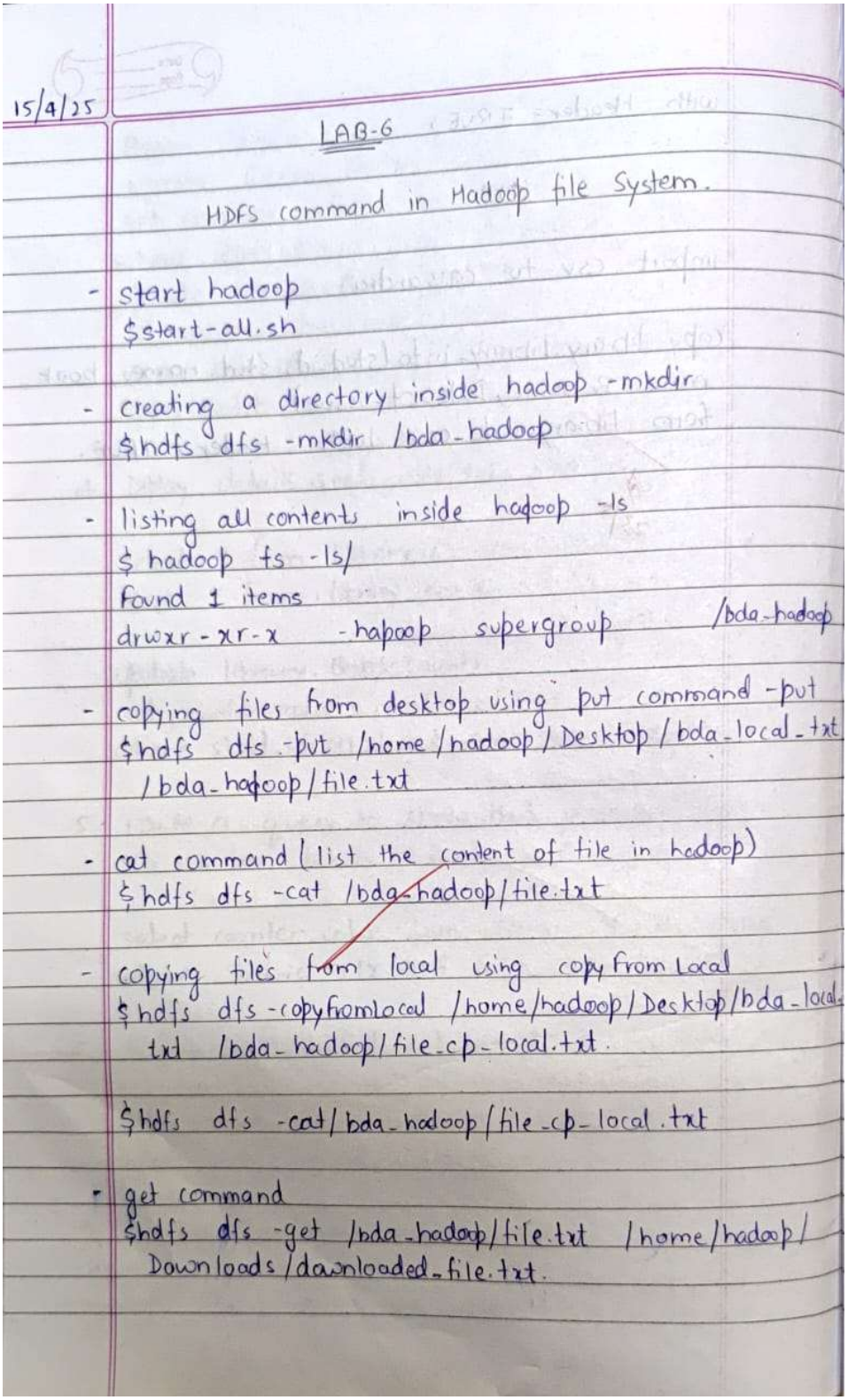
```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.1.0 | Cassandra 4.1.8 | CQL spec 3.4.6 | Native protocol v5]
Use HELP for help.
cqlsh> create keyspace library with replication={'class':'Simplestrategy','replication_factor':1};
ConfigurationException: Unable to find replication strategy class 'org.apache.cassandra.locator.Simplestrategy'
cqlsh> create keyspace library with replication={'class':'Simplestrategy','replication_factor':1};exit
ConfigurationException: Unable to find replication strategy class 'org.apache.cassandra.locator.Simplestrategy'
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.1.0 | Cassandra 4.1.8 | CQL spec 3.4.6 | Native protocol v5]
Use HELP for help.
cqlsh> create keyspace library with replication={'class':'SimpleStrategy','replication_factor':1};exit
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.1.0 | Cassandra 4.1.8 | CQL spec 3.4.6 | Native protocol v5]
Use HELP for help.
cqlsh> create keyspace library with replication={'class':'SimpleStrategy','replication_factor':1};
AlreadyExists: Keyspace 'library' already exists
cqlsh> exit
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ cqlsh
Connected to Test Cluster at 127.0.0.1:9042
[cqlsh 6.1.0 | Cassandra 4.1.8 | CQL spec 3.4.6 | Native protocol v5]
Use HELP for help.
cqlsh> create keyspace libraries with replication={'class':'SimpleStrategy','replication_factor':1};
cqlsh> keyspaces
   ...
cqlsh> describe keyspaces;

libraries  students     system_distributed  system_views
library    system       system_schema       system_virtual_schema
student    system_auth  system_traces

cqlsh> use libraries;
cqlsh:libraries> create table l_info(sid int primary key, c_val counter, sname varchar,bname varchar,bid int,doi timestamp);
InvalidRequest: Error from server: code=2200 [Invalid query] message="Cannot mix counter and non counter columns in the same tabl
e"
cqlsh:libraries> create table l_info(sid int primary key, sname varchar,bname varchar,bid int,doi timestamp);
cqlsh:libraries> create table count(sid int primary key,c_val counter);
cqlsh:libraries> begin batch
            ... insert into l_info(sid,sname,bname,bid,doi)
            ... values(112,'alice','bda',1,'2020-03-03')
            ... insert into l_info(sid,sname,bname,bid,doi)
            ... values(113,'preeti','cn',2,'2020-03-04')
            ... apply batch;
cqlsh:libraries> update l_info
```

```
cqlsh:libraries> select * from l_info;

 sid | bid | bname | doi                                 | sname
-----+-----+-------+-------------------------------------+--------
 113 |   2 |    cn | 2020-03-03 18:30:00.000000+0000 | preeti
 112 |   1 |   bda | 2020-03-02 18:30:00.000000+0000 |  alice

(2 rows)
cqlsh:libraries> select * from count;

 sid | c_val
-----+-------
 112 |     1

(1 rows)
cqlsh:libraries>
```

Program 4

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

Observation:



15/4/25

LAB-6

HDFS command in Hadoop file System.

- start hadoop
  $start-all.sh
- creating a directory inside hadoop -mkdir
  $hdfs dfs -mkdir /bda_hadoop
- listing all contents inside hadoop -ls
  $ hadoop fs -ls/
  Found 1 items
  drwxr-xr-x - hapoop supergroup /bda_hadoop
- copying files from desktop using put command -put
  $hdfs dfs -put /home/hadoop/Desktop/bda_local_txt /bda_hadoop/file.txt
- cat command (list the content of file in hadoop)
  $hdfs dfs -cat /bda_hadoop/file.txt
- copying files from local using copy from Local
  $hdfs dfs -copyfromlocal /home/hadoop/Desktop/bda_local.txt /bda_hadoop/file_cp_local.txt.

  $hdfs dfs -cat/bda_hadoop/file_cp_local.txt
- get command
  $hdfs dfs -get /bda_hadoop/file.txt /home/hadoop/Downloads/downloaded_file.txt.

$ hdfs dfs -getmerge /bda_hadoop/ /home/hadoop/Downloads/merged_output.txt.

- display Access Control List (ACL) permission of file or directory in HDFS.
$ hadoop fs -getfacl /bda_hadoop/

- copyToLocal
$ hdfs dfs -copyToLocal /bda_hadoop/file.txt /home/hadoop/Desktop

- mv
$ hadoop fs -mv /bda_hadoop /abc
$ hadoop fs -ls /abc

- copy
$ hadoop fs -cp /hello/ /hadoop_lab

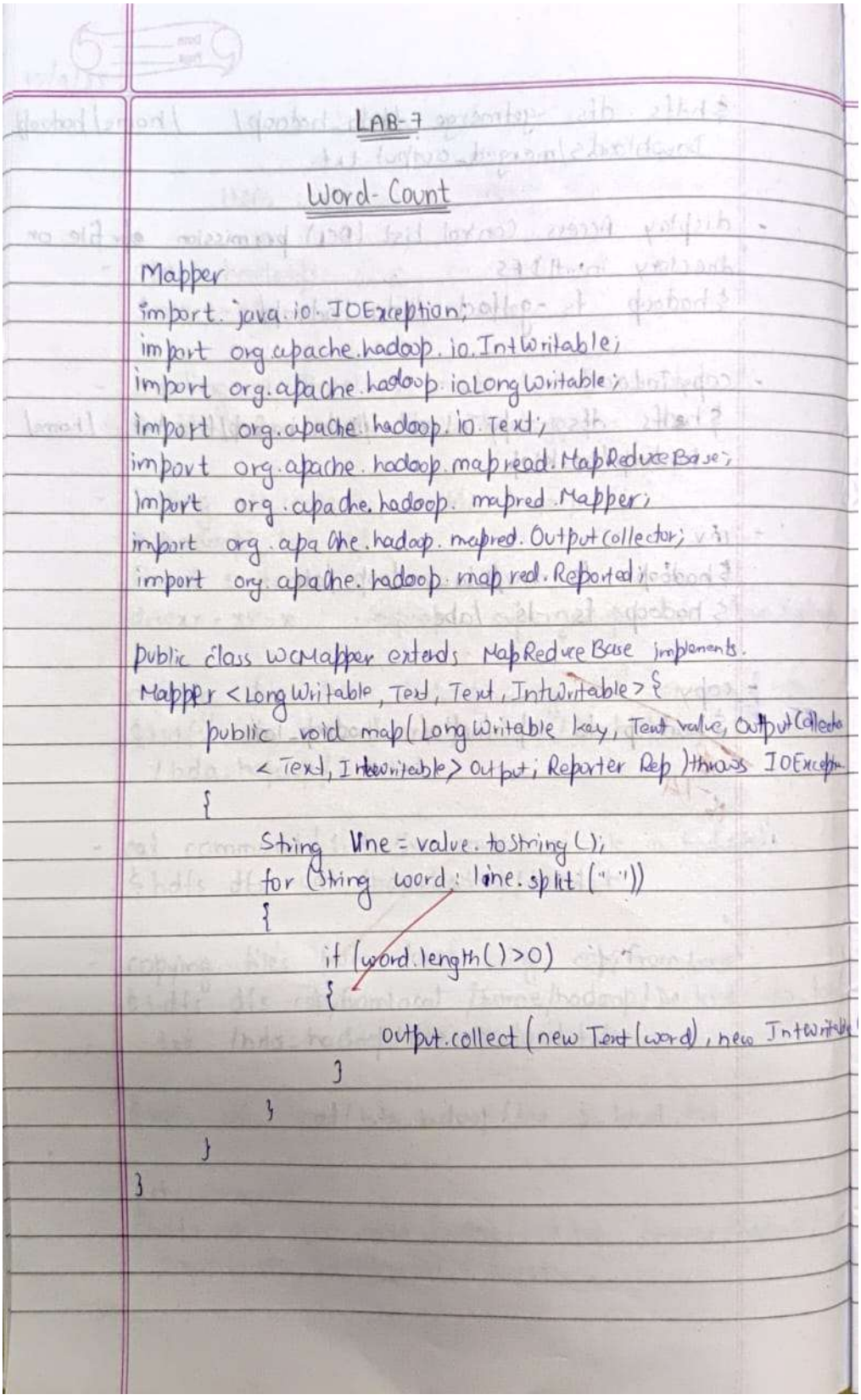
Code with Output:

```
hadoop@bmscecse-HP-Elite-Tower-600-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 [bmscecse-HP-Elite-Tower-600-G9-Desktop-PC]
Starting resourcemanager
Starting nodemanagers
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -mkdir /bda_hadoop
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -ls /
Found 4 items
drwxr-xr-x   - hadoop supergroup          0 2025-04-15 15:07 /abc
drwxr-xr-x   - hadoop supergroup          0 2025-05-26 14:13 /bda_hadoop
drwxr-xr-x   - hadoop supergroup          0 2025-05-22 16:32 /pqr
drwxr-xr-x   - hadoop supergroup          0 2025-05-20 16:36 /rgs
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -put /home/hadoop/sample.txt /bda_hadoop/file.txt
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -copyFromLocal /home/hadoop/sample.txt /bda_hadoop/local.txt
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -cat /bda_hadoop/file.txt
hi how are you
how is your job
how is your family
how is your brother
how is your sister
eof
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -get /bda_hadoop/file.txt /home/hadoop/sample.txt
get: `/home/hadoop/sample.txt': File exists
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -get /bda_hadoop/file.txt /home/hadoop/get.txt
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop fs -getfacl /bda_hadoop/
# file: /bda_hadoop
# owner: hadoop
# group: supergroup
user::rwx
group::r-x
other::r-x

hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -copyToLocal /bda_hadoop/file.txt /home/hadoop/tolocal.txt
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop fs -cp /bda_hadoop /abc
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop fs -ls /abc
Found 3 items
drwxr-xr-x   - hadoop supergroup          0 2025-05-26 14:28 /abc/bda_hadoop
-rw-r--r--   1 hadoop supergroup         55 2025-04-15 15:05 /abc/file.txt
-rw-r--r--   1 hadoop supergroup         55 2025-04-15 15:07 /abc/file_cp_.txt
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$
```

Program 5

Implement Wordcount program on Hadoop framework

Observation:



LAB-7

Word-Count

Mapper

```
import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.MapReduceBase;
import org.apache.hadoop.mapred.Mapper;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reporter;

public class WcMapper extends MapReduceBase implements
Mapper<LongWritable, Text, Text, IntWritable>{
    public void map(LongWritable key, Text value, OutputCollecto
        <Text, IntWritable> output, Reporter Rep) throws IOExcepti
    {
        String line = value.toString();
        for (String word : line.split(" "))
        {
            if (word.length()>0)
            {
                output.collect(new Text(word), new IntWritable
            }
        }
    }
}
```

WC Reducer.

```
import java.io.IOExcepher;
import java.util.Iterator;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapred.Reporter;

public class WCReducer extends MapReduceBase implement
Reducer<Text, IntWritable, Text, IntWritable>{
    public void reduce(Text key, Iterator<IntWritable>
      value, OutputCollector<Text, IntWritable>outpt,
       Reporter rep) throws IOException
    {
        int count=0;
        while(value.hasNext())
        {
            intWritable i=value.next();
            count+=i.get();
        }
        output.collect(key, new IntWritable(count));
    }
}
```

WCDriver

```
import java.io.IOException;
import org.apache.hadoop.conf.configure;
                    fs.path;
                    io.IntWritable;
                    io.Text;
                    mapred.FileInputFormat;
                    mapred.FileOutputFormat;
                    mapred.JobClient;
                    mapred.JobConf;
                    util.Tool;
                    util.ToolRunner;

public class WCDriver extends Configured implements Tool
    public int run(String args[]) throws IOException
    {
        if (args.length < 2)
        {
            System.out.println("Please give valid inputs");
            return -1;
        }
        JobConf conf = new JobConf(WCDriver.class);
        FileInputFormat.setInputPaths(conf, new Path(args[0]));
        FileOutputFormat.setOutputPath(conf, new Path(args[1]));
        conf.setMapperClass(WCMapper.class);
        conf.setReducerClass(WCReducer.class);
        conf.setMapOutputKeyClass(Text.class);
        conf.setMapOutputValueClass(IntWritable.class);
        conf.setOutputKeyClass(Text.class);
        conf.setOutputValueClass(IntWritable.class);
        JobClient.runJob(conf);
        return 0;
    }
```

```
public static void main (String args[]) throws Exception.
{
    int exitCode = ToolRunner.run (new WCDriver (), args);
    System.out.println (exitCode);
}
}
```

23/5

Code with Output:

```
hadoop@bmscecse-HP-Elite-Tower-600-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]
localhost: namenode is running as process 12082.  Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 12255.  Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscecse-HP-Elite-Tower-600-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-600-G9-Desktop-PC: secondarynamenode is running as process 12557.  Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
Starting resourcemanager
resourcemanager is running as process 12845.  Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 13014.  Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ jps
12082 NameNode
13014 NodeManager
15100 org.eclipse.equinox.launcher_1.6.1000.v20250227-1734.jar
17036 Jps
12845 ResourceManager
12557 SecondaryNameNode
12255 DataNode
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -copyFromLocal /home/hadoop/sample.txt /bda_hadoop/input.txt
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/hdpwordcount.jar WCDriver /bda_hadoop/input.txt /bda_hadoop/output
Exception in thread "main" java.lang.ClassNotFoundException: WCDriver
        at java.base/java.net.URLClassLoader.findClass(URLClassLoader.java:476)
        at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:594)
        at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:527)
        at java.base/java.lang.Class.forName0(Native Method)
        at java.base/java.lang.Class.forName(Class.java:398)
        at org.apache.hadoop.util.RunJar.run(RunJar.java:321)
        at org.apache.hadoop.util.RunJar.main(RunJar.java:241)
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/hdpwordcount.jar hdpwordcount.WCDriver /bda_hadoop/input.txt /bda_hadoop/output
2025-05-26 14:40:01,404 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-05-26 14:40:01,440 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-05-26 14:40:01,440 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-05-26 14:40:01,446 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-05-26 14:40:01,501 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-05-26 14:40:01,545 INFO mapred.FileInputFormat: Total input files to process : 1
2025-05-26 14:40:01,567 INFO mapreduce.JobSubmitter: number of splits:1
2025-05-26 14:40:01,624 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local276129153_0001
2025-05-26 14:40:01,624 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-26 14:40:01,677 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-05-26 14:40:01,679 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-05-26 14:40:01,679 INFO mapreduce.Job: Running job: job_local276129153_0001
2025-05-26 14:40:01,680 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2025-05-26 14:40:01,682 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-26 14:40:01,682 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ign
```

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -ls /bda_hadoop/output
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2025-05-26 14:40 /bda_hadoop/output/_SUCCESS
-rw-r--r--   1 hadoop supergroup         75 2025-05-26 14:40 /bda_hadoop/output/part-00000
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -cat /bda_hadoop/output/part-00000
are     1
brother 1
eof     1
family  1
hi      1
how     5
is      4
job     1
sister  1
you     1
your    4
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$
```

Program 6

From the following link extract the weather data

https://github.com/tomwhite/hadoop-book/tree/master/input/ncdc/all

a) Create a MapReduce program to find average temperature for each year from the NCDC data set.

b) find the mean max temperature for every month

Observation:

Code with Output:

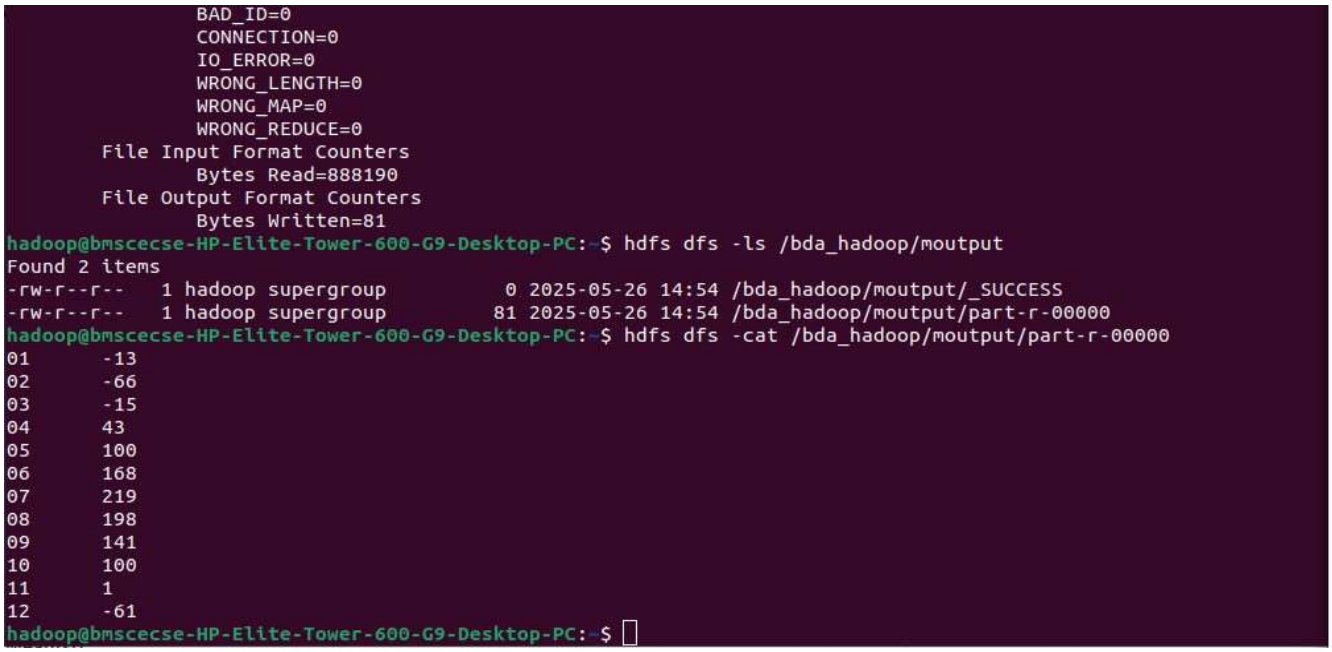
a) Average temperature

```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ jps
12082 NameNode
17908 Jps
13014 NodeManager
15100 org.eclipse.equinox.launcher_1.6.1000.v20250227-1734.jar
12845 ResourceManager
12557 SecondaryNameNode
12255 DataNode
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -copyFromLocal /home/hadoop/Downloads/1901 /bda_hadoop/avinput.txt
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/WeatherAverage.jar WeatherAverage.AVDriver /bda_h
adoop/avinput.txt /bda_hadoop/avoutput
2025-05-26 14:49:09,290 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-05-26 14:49:09,327 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-05-26 14:49:09,327 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-05-26 14:49:09,380 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool inte
rface and execute your application with ToolRunner to remedy this.
2025-05-26 14:49:09,427 INFO input.FileInputFormat: Total input files to process : 1
2025-05-26 14:49:09,452 INFO mapreduce.JobSubmitter: number of splits:1
2025-05-26 14:49:09,510 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1313646497_0001
2025-05-26 14:49:09,510 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-26 14:49:09,566 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-05-26 14:49:09,566 INFO mapreduce.Job: Running job: job_local1313646497_0001
2025-05-26 14:49:09,567 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-05-26 14:49:09,570 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitte
rFactory
2025-05-26 14:49:09,571 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-26 14:49:09,571 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:f
alse, ignore cleanup failures: false
2025-05-26 14:49:09,571 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2025-05-26 14:49:09,620 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-05-26 14:49:09,620 INFO mapred.LocalJobRunner: Starting task: attempt_local1313646497_0001_m_000000_0
2025-05-26 14:49:09,629 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitte
rFactory
2025-05-26 14:49:09,629 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-26 14:49:09,629 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:f
alse, ignore cleanup failures: false
2025-05-26 14:49:09,635 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-05-26 14:49:09,637 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/bda_hadoop/avinput.txt:0+888190
2025-05-26 14:49:09,666 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-05-26 14:49:09,666 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-05-26 14:49:09,666 INFO mapred.MapTask: soft limit at 83886080
2025-05-26 14:49:09,666 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-05-26 14:49:09,666 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-05-26 14:49:09,668 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-26 14:49:09,730 INFO mapred.LocalJobRunner:
2025-05-26 14:49:09,731 INFO mapred.MapTask: Starting flush of map output
2025-05-26 14:49:09,731 INFO mapred.MapTask: Spilling map output
2025-05-26 14:49:09,731 INFO mapred.MapTask: bufstart = 0; bufend = 59076; bufvoid = 104857600
2025-05-26 14:49:09,731 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26188144(104752576); length = 26253/6553600
2025-05-26 14:49:09,739 INFO mapred.MapTask: Finished spill 0
2025-05-26 14:49:09,743 INFO mapred.Task: Task:attempt_local1313646497_0001_m_000000_0 is done. And is in the process of committing
2025-05-26 14:49:09,745 INFO mapred.LocalJobRunner: map
```

```
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=1052770304
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=888190
        File Output Format Counters
                Bytes Written=8
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -ls /bda_hadoop/avoutput
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2025-05-26 14:49 /bda_hadoop/avoutput/_SUCCESS
-rw-r--r--   1 hadoop supergroup          8 2025-05-26 14:49 /bda_hadoop/avoutput/part-r-00000
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -cat /bda_hadoop/avoutput/part-r-00000
1901    46
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$
```

b) Maximum temperature

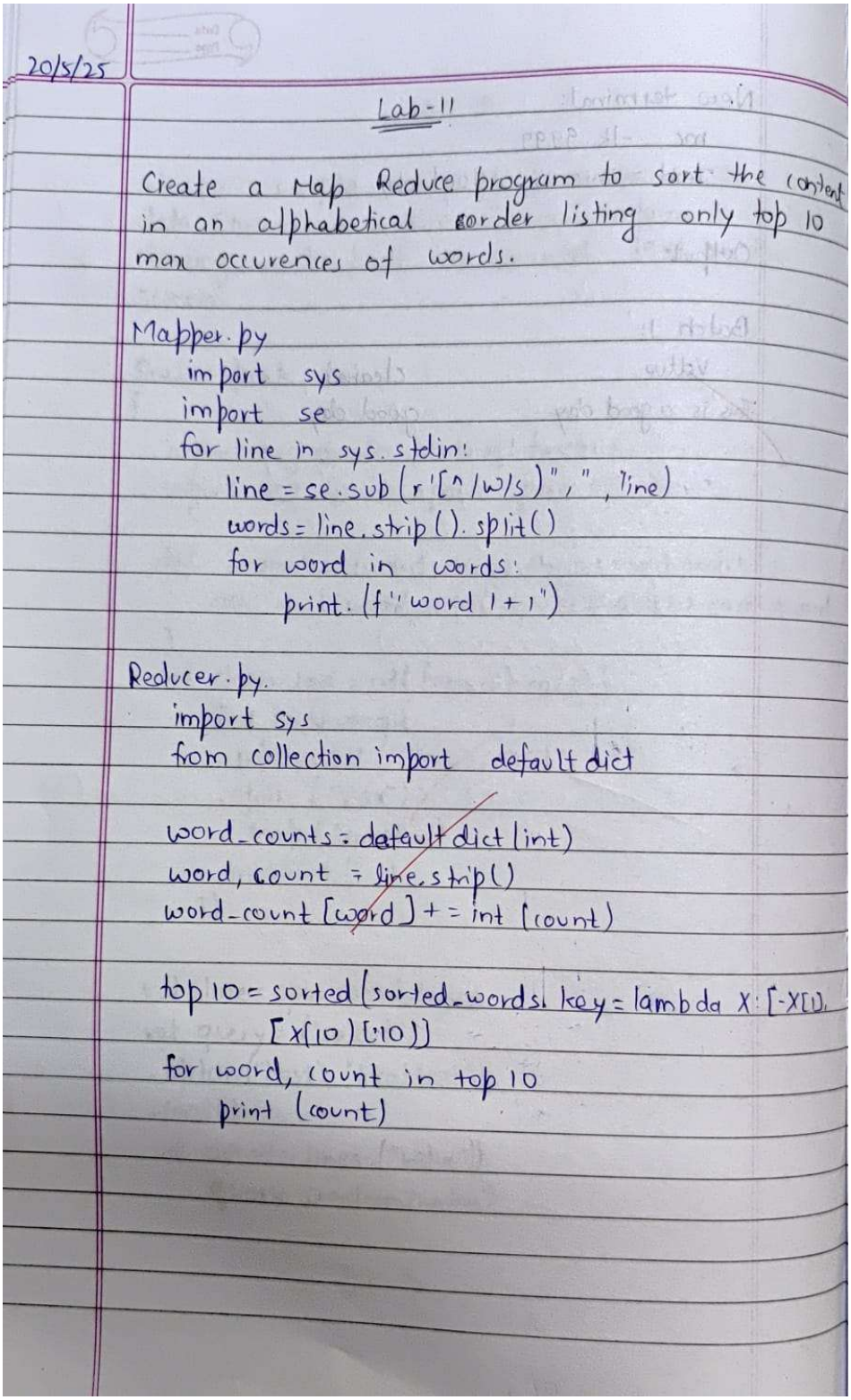
```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ jps
18721 Jps
12082 NameNode
13014 NodeManager
15100 org.eclipse.equinox.launcher_1.6.1000.v20250227-1734.jar
12845 ResourceManager
12557 SecondaryNameNode
12255 DataNode
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -copyFromLocal /home/hadoop/Downloads/1901 /bda_ha
doop/minput.txt
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/meanTemp.jar Mean.MNDriver
/bda_hadoop/minput.txt /bda_hadoop/moutput
2025-05-26 14:54:41,993 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-05-26 14:54:42,029 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-05-26 14:54:42,029 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-05-26 14:54:42,083 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. I
mplement the Tool interface and execute your application with ToolRunner to remedy this.
2025-05-26 14:54:42,131 INFO input.FileInputFormat: Total input files to process : 1
2025-05-26 14:54:42,158 INFO mapreduce.JobSubmitter: number of splits:1
2025-05-26 14:54:42,216 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local862196817_0001
2025-05-26 14:54:42,216 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-26 14:54:42,272 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-05-26 14:54:42,273 INFO mapreduce.Job: Running job: job_local862196817_0001
2025-05-26 14:54:42,273 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-05-26 14:54:42,276 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting
 to FileOutputCommitterFactory
2025-05-26 14:54:42,277 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-26 14:54:42,277 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders un
der output directory:false, ignore cleanup failures: false
2025-05-26 14:54:42,277 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.F
ileOutputCommitter
2025-05-26 14:54:42,319 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-05-26 14:54:42,319 INFO mapred.LocalJobRunner: Starting task: attempt_local862196817_0001_m_000000_0
2025-05-26 14:54:42,328 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting
 to FileOutputCommitterFactory
2025-05-26 14:54:42,329 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-26 14:54:42,329 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders un
der output directory:false, ignore cleanup failures: false
2025-05-26 14:54:42,335 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-05-26 14:54:42,336 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/bda_hadoop/minput.txt:0+88
8190
2025-05-26 14:54:42,366 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-05-26 14:54:42,366 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-05-26 14:54:42,366 INFO mapred.MapTask: soft limit at 83886080
2025-05-26 14:54:42,366 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-05-26 14:54:42,366 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-05-26 14:54:42,368 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapO
utputBuffer
2025-05-26 14:54:42,428 INFO mapred.LocalJobRunner:
2025-05-26 14:54:42,428 INFO mapred.MapTask: Starting flush of map output
2025-05-26 14:54:42,429 INFO mapred.MapTask: Spilling map output
2025-05-26 14:54:42,429 INFO mapred.MapTask: bufstart = 0; bufend = 45948; bufvoid = 104857600
```


BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=888190
File Output Format Counters
Bytes Written=81
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -ls /bda_hadoop/moutput
Found 2 items
-rw-r--r-- 1 hadoop supergroup 0 2025-05-26 14:54 /bda_hadoop/moutput/_SUCCESS
-rw-r--r-- 1 hadoop supergroup 81 2025-05-26 14:54 /bda_hadoop/moutput/part-r-00000
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -cat /bda_hadoop/moutput/part-r-00000
01 -13
02 -66
03 -15
04 43
05 100
06 168
07 219
08 198
09 141
10 100
11 1
12 -61
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$

Program 7

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

Observation:



20/5/25

Lab-11

Create a Map Reduce program to sort the content in an alphabetical order listing only top 10 max occurences of words.

Mapper.py

```
import sys
import re
for line in sys.stdin:
    line = re.sub(r'[^\w\s]', '', line)
    words = line.strip().split()
    for word in words:
        print(f"{word}\t1")
```

Reducer.py

```
import sys
from collection import defaultdict

word_counts = defaultdict(int)
word, count = line.strip()
word_count[word] += int(count)

top10 = sorted(sorted_words, key = lambda x: [-x[1],
        [x[10][:10]])
for word, count in top10
    print(count)
```

Output:
apple : 34
code : 25
dots : 23
file : 21
hadoop: 20
text : 17
zebra : 25

23/5/25

Code with Output:

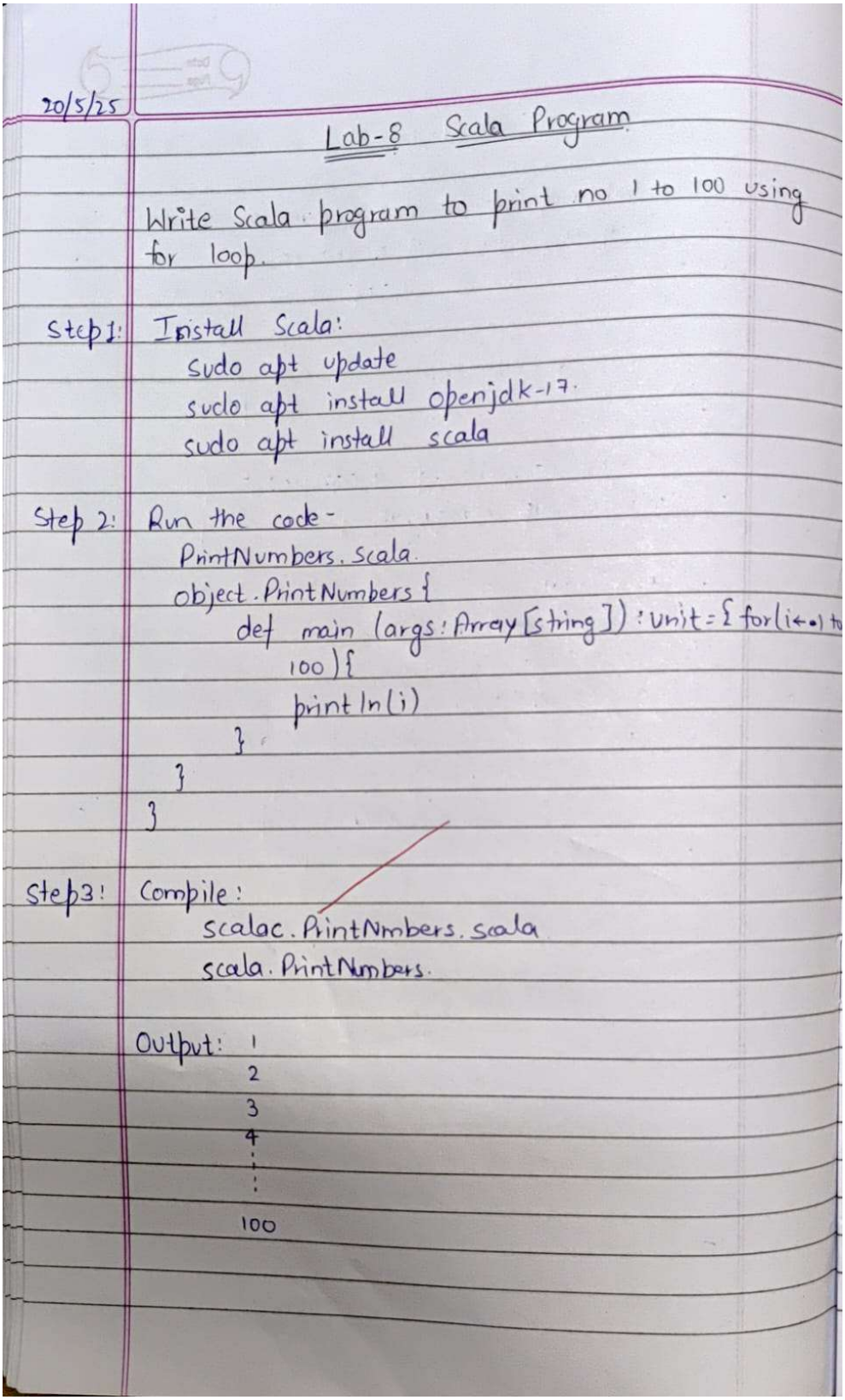
```
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ jps
12082 NameNode
19238 Jps
13014 NodeManager
15100 org.eclipse.equinox.launcher_1.6.1000.v20250227-1734.jar
12845 ResourceManager
12557 SecondaryNameNode
12255 DataNode
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -copyFromLocal /home/hadoop/Desktop/TopN.txt /bda_
haoop/tinput.txt
copyFromLocal: `/bda_haoop/tinput.txt': No such file or directory: `hdfs://localhost:9000/bda_haoop/tinput.txt'
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -copyFromLocal /home/hadoop/Desktop/TopN.txt /bda_
hadoop/tinput.txt
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/TopN.jar TopN.TNDriver /bda
_hadoop/tinput.txt /bda_hadoop/toutput
2025-05-26 14:59:03,334 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-05-26 14:59:03,372 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-05-26 14:59:03,372 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-05-26 14:59:03,426 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. I
mplement the Tool interface and execute your application with ToolRunner to remedy this.
2025-05-26 14:59:03,472 INFO input.FileInputFormat: Total input files to process : 1
2025-05-26 14:59:03,497 INFO mapreduce.JobSubmitter: number of splits:1
2025-05-26 14:59:03,554 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1824101299_0001
2025-05-26 14:59:03,554 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-26 14:59:03,609 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-05-26 14:59:03,610 INFO mapreduce.Job: Running job: job_local1824101299_0001
2025-05-26 14:59:03,610 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-05-26 14:59:03,614 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting
 to FileOutputCommitterFactory
2025-05-26 14:59:03,614 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-26 14:59:03,614 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders un
der output directory:false, ignore cleanup failures: false
2025-05-26 14:59:03,614 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.F
ileOutputCommitter
2025-05-26 14:59:03,654 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-05-26 14:59:03,655 INFO mapred.LocalJobRunner: Starting task: attempt_local1824101299_0001_m_000000_0
2025-05-26 14:59:03,664 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting
 to FileOutputCommitterFactory
2025-05-26 14:59:03,664 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-26 14:59:03,664 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders un
der output directory:false, ignore cleanup failures: false
2025-05-26 14:59:03,670 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-05-26 14:59:03,672 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/bda_hadoop/tinput.txt:0+95
2025-05-26 14:59:03,701 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-05-26 14:59:03,701 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-05-26 14:59:03,701 INFO mapred.MapTask: soft limit at 83886080
2025-05-26 14:59:03,701 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-05-26 14:59:03,701 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-05-26 14:59:03,702 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapO
utputBuffer
2025-05-26 14:59:03,738 INFO mapred.LocalJobRunner:
2025-05-26 14:59:03,739 INFO mapred.MapTask: Starting flush of map output
```

```
        File System Counters
                FILE: Number of bytes read=10682
                FILE: Number of bytes written=1291808
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=190
                HDFS: Number of bytes written=40
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=3
                Map output records=15
                Map output bytes=154
                Map output materialized bytes=190
                Input split bytes=108
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=190
                Reduce input records=15
                Reduce output records=5
                Spilled Records=30
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=1052770304
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=95
        File Output Format Counters
                Bytes Written=40
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -ls /bda_hadoop/toutput
Found 2 items
-rw-r--r--   1 hadoop supergroup          0 2025-05-26 14:59 /bda_hadoop/toutput/_SUCCESS
-rw-r--r--   1 hadoop supergroup         40 2025-05-26 14:59 /bda_hadoop/toutput/part-r-00000
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ hdfs dfs -cat /bda_hadoop/toutput/part-r-00000
banana  5
apple   4
fruit   3
mango   2
kiwi    1
hadoop@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$
```

Program 8

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

Observation:



20/5/25

Lab-8 Scala Program

Write Scala program to print no 1 to 100 using for loop.

Step 1: Install Scala:

```
sudo apt update
sudo apt install openjdk-17.
sudo apt install scala
```

Step 2: Run the code -

PrintNumbers.scala

```
object PrintNumbers {
    def main (args: Array[String]): Unit = { for(i<-1 to
        100){
        println(i)
    }
  }
}
```

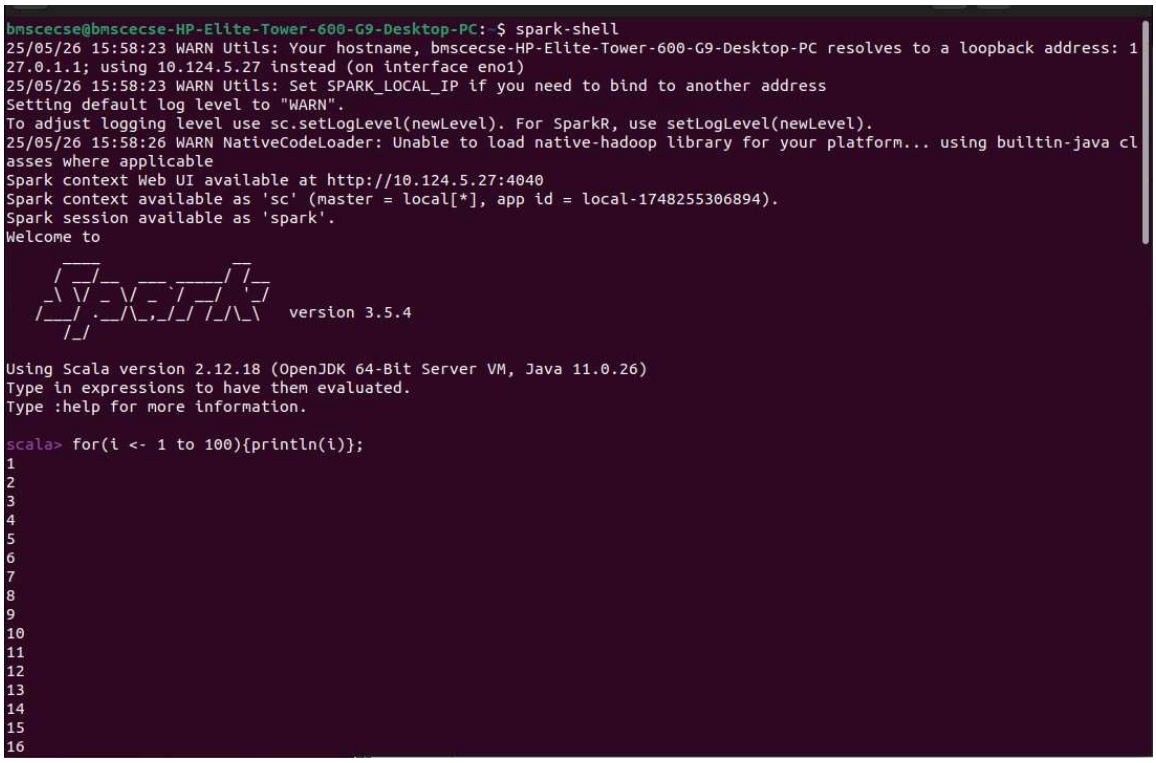
Step 3: Compile:

```
scalac PrintNmbers.scala
scala PrintNumbers
```

Output:

```
1
2
3
4
:
:
100
```

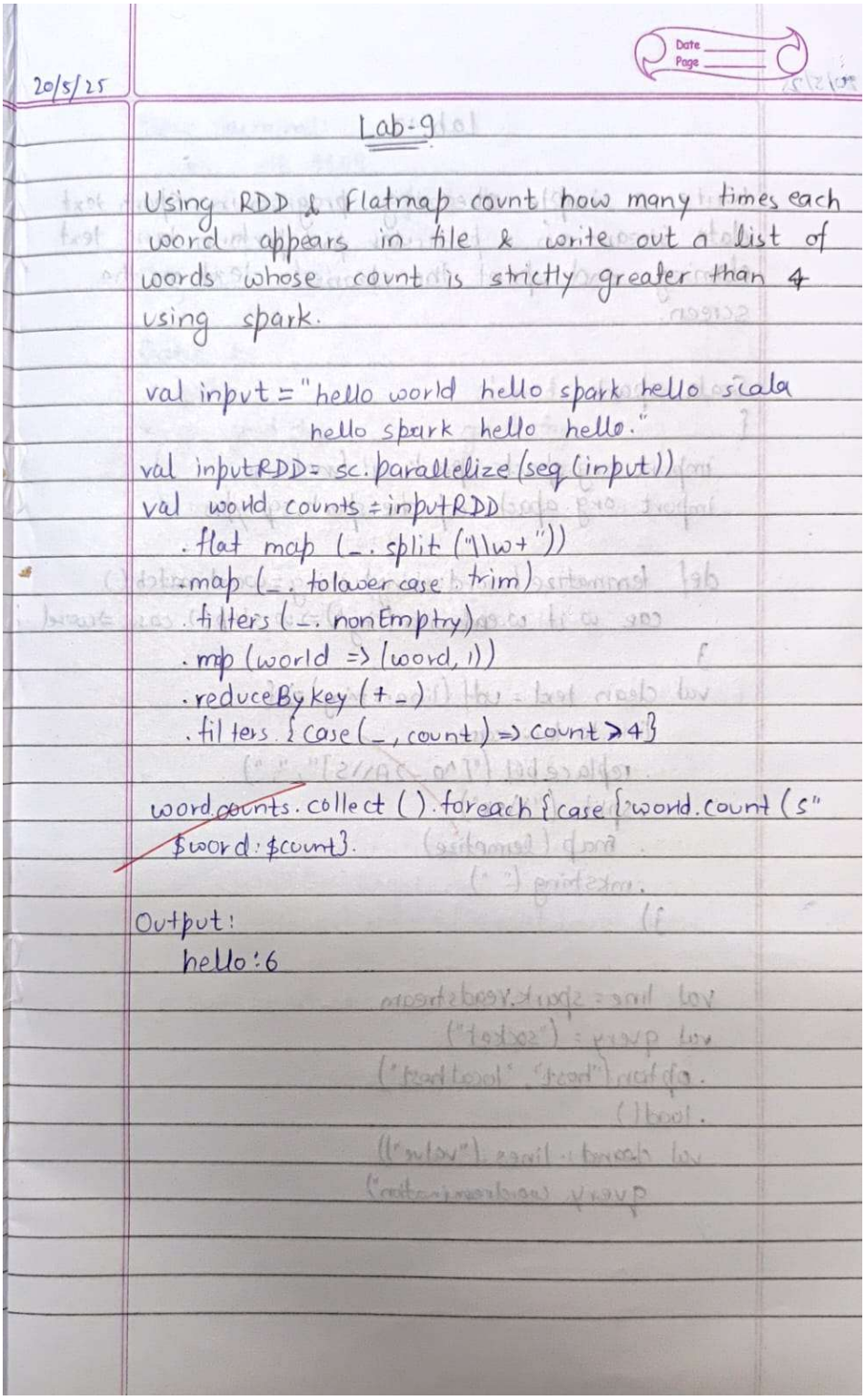
Code with Output:



Program 9

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

Observation:



20/5/25

Lab-9

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

```
val input = "hello world hello spark hello scala
             hello spark hello hello."
val inputRDD = sc.parallelize(seq(input))
val word counts = inputRDD
  .flat map (_.split("\\w+"))
  .map(_.tolowercase.trim)
  .filters(_.nonEmptry)
  .mp(world => (word, 1))
  .reduceByKey(+_)
  .filters.{case(_, count) => count > 4}

word counts.collect().foreach{case{(word.count(s"
$word: $count}.
```

Output:

hello: 6

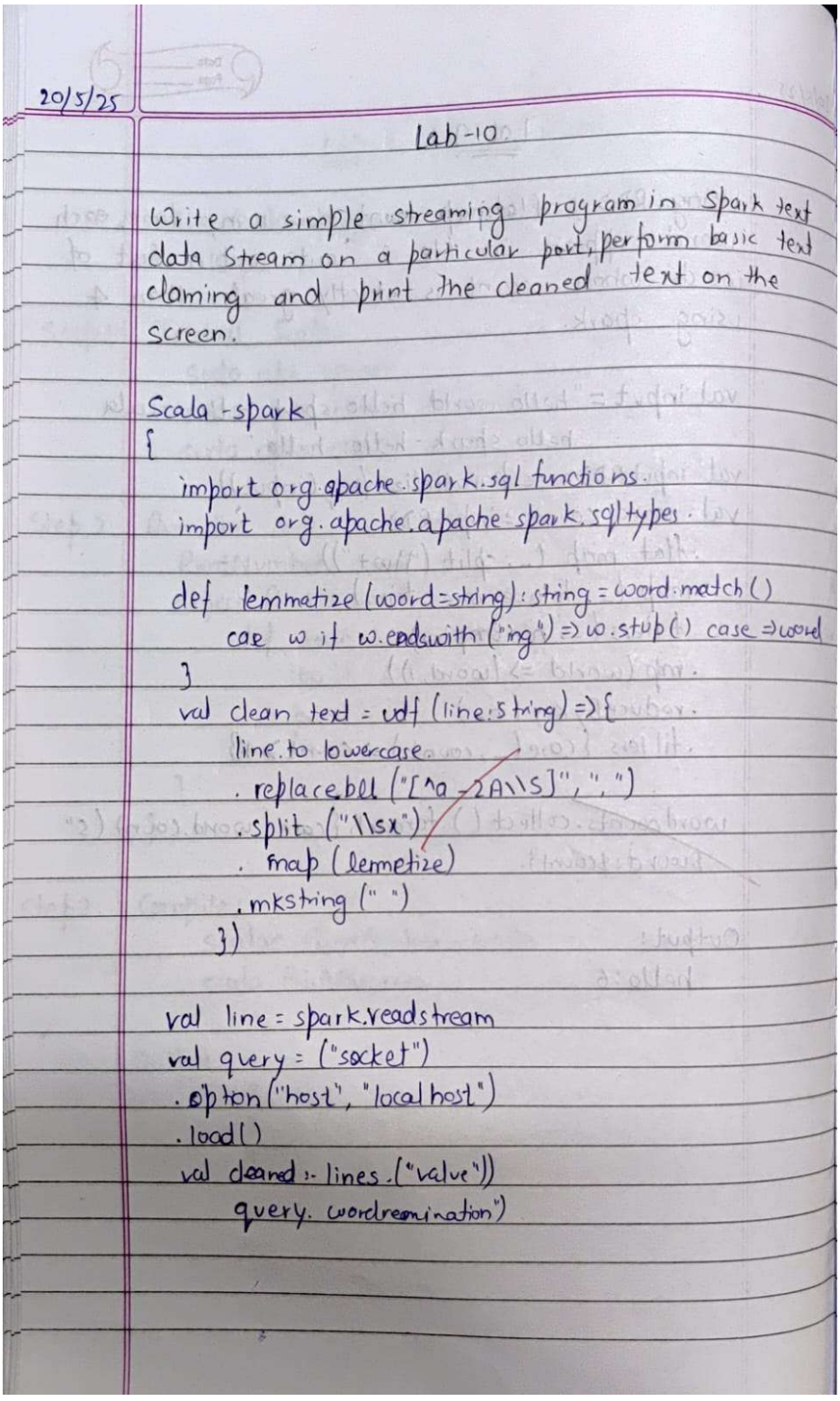
Code with Output:

```
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ echo "code code code code code spark spark spark spark spark hell
o hello hi hi joe ken">input.txt
bmscecse@bmscecse-HP-Elite-Tower-600-G9-Desktop-PC:~$ spark-shell
25/05/26 16:01:15 WARN Utils: Your hostname, bmscecse-HP-Elite-Tower-600-G9-Desktop-PC resolves to a loopback address:
127.0.1.1; using 10.124.5.27 instead (on interface eno1)
25/05/26 16:01:15 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
25/05/26 16:01:17 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java c
lasses where applicable
Spark context Web UI available at http://10.124.5.27:4040
Spark context available as 'sc' (master = local[*], app id = local-1748255477930).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.5.4
      /_/

Using Scala version 2.12.18 (OpenJDK 64-Bit Server VM, Java 11.0.26)
Type in expressions to have them evaluated.
Type :help for more information.

scala> val lines=sc.textFile("input.txt")
lines: org.apache.spark.rdd.RDD[String] = input.txt MapPartitionsRDD[1] at textFile at <console>:23

scala> val words=lines.flatmap(line => line.split(" "))
<console>:23: error: value flatmap is not a member of org.apache.spark.rdd.RDD[String]
       val words=lines.flatmap(line => line.split(" "))
                       ^

scala> val words=lines.flatMap(line => line.split(" "))
words: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at flatMap at <console>:23

scala> val wordParts = words.map(word => (word,1))
wordParts: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[3] at map at <console>:23

scala> val wordcount = wordParts.reduceByKey(_+_)
wordcount: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[4] at reduceByKey at <console>:23

scala> val freq = wordcount.filter {case (word,count) => count > 4}
freq: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[5] at filter at <console>:23

scala> freq.collect().foreach(println)
(spark,5)
(code,5)

scala>
```

Program 10

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. (Open Ended Question).

Observation:



20/5/25

Lab-10

Write a simple streaming program in Spark text data stream on a particular port, perform basic text claming and print the cleaned text on the screen.

```
Scala -spark
{
  import org.apache.spark.sql functions.
  import org.apache.apache spark.sql types.

  def lemmatize (word=string): string = word.match()
     case w if w.endswith ("ing") => w.stup() case => word
  }
  val clean text = udf (line:string) => {
     line.to lowercase
     . replaceAll ("[^a-zA\\s]", "")
     .split ("\\sx")
     . map (lemetize)
     .mkstring (" ")
    })

  val line = spark.readstream
  val query = ("socket")
  .option("host", "localhost")
  .load()
  val cleaned :- lines.("value"))
     query. wordreamination")
```

New terminal:

mc -lk.9999

Text = This is a good day.

Output:

Batch 1:

Value	cleaned
This is a good day	good day.