VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"JnanaSangama", Belgaum -590014, Karnataka.

LAB REPORT
on

Big Data Analytics(23CS6PCBDA)

Submitted by

Prabhanjan Bhat(1BM22CS196)

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 **Prabhanjan Bhat(1BM22CS196),** 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 2024. The Lab report has been approved as it satisfies the academic requirements in respect of a **Big Data Analytics - (23CS6PCBDA)** work prescribed for the said degree.

Ms. Ambuja K
Assistant Professor
Department of CSE
BMSCE, Bengaluru

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

Index Sheet

	NCDC data set. b) find the mean max temperature for every month.	
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.	39
8	Write a Scala program to print numbers from 1 to 100 using for loop.	46
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.	49
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).	50

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:

```
Atlas atlas-wanmtx-shard-0 [primary] Student> use Students
switched to db Students
Atlas atlas-wanmtx-shard-0 [primary] Students> show collections

Atlas atlas-wanmtx-shard-0 [primary] Students> db.students.insertMany([
...     { "Rollno": 10, "Name": "John", "Age": 20, "ContactNo": "1234567890", "Email-Id":
"john@example.com", "grade": "A", "hobby": "Reading" },
...     { "Rollno": 11, "Name": "Alice", "Age": 21, "ContactNo": "9876543210", "Email-Id":
 "alice@example.com", "grade":
 "B", "hobby": "Painting" },
...     { "Rollno": 12, "Name": "Bob", "Age": 22, "ContactNo": "2345678901", "Email-Id": "
bob@example.com", "grade": "C", "hobby": "Cooking" },
...     { "Rollno": 13, "Name": "Eve", "Age": 23, "ContactNo": "3456789012", "Email-Id": "
eve@example.com", "grade": "A"
 },
...     { "Rollno": 14, "Name": "Charlie", "Age": 24, "ContactNo": "4567890123", "Email-Id
": "charlie@example.com", "hobby": "Gardening" }
... ])
{
  acknowledged: true,
  insertedIds: {
    '0': ObjectId("661ce9dc76a00ff8cc51dae1"),
    '1': ObjectId("661ce9dc76a00ff8cc51dae2"),
    '2': ObjectId("661ce9dc76a00ff8cc51dae3"),
    '3': ObjectId("661ce9dc76a00ff8cc51dae4"),
    '4': ObjectId("661ce9dc76a00ff8cc51dae5")
  }
}
```

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

db.students.updateOne(

{ "Rollno": 10 },

{ $set: { "Email-Id": "john.doe@example.com" } }

)

```
Atlas atlas-wanmtx-shard-0 [primary] Students> db.students.updateOne(
...     { "Rollno": 10 },
...     { $set: { "Email-Id": "john.doe@example.com" } }
... )
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
```

2. **Replace the student name from “Alice” to “Alicee” of rollno 11**

db.students.updateOne(

{ "Rollno": 11 },

{ $set: { "Name": "Alicee" } }

)

```
Atlas atlas-wanmtx-shard-0 [primary] Students> db.students.updateOne(
...     { "Rollno": 11 },
...     { $set: { "Name": "Alicee" } }
... )
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
```

3. **Display Student Name and grade(Add if grade is not present)where the _id column is 1.**

db.students.find({}, { "Name": 1, "grade": { $ifNull: ["$grade", "Not available"] }, "_id": 0 })

```
Atlas atlas-wanmtx-shard-0 [primary] Students> db.students.find({}, { "Name": 1, "grade":
{ $ifNull: ["$grade", "Not available"] }, "_id": 0 })
[
  { Name: 'John', grade: 'A' },
  { Name: 'Alicee', grade: 'B' },
  { Name: 'Bob', grade: 'C' },
  { Name: 'Eve', grade: 'A' },
  { Name: 'Charlie', grade: 'Not available' }
]
```

4. **Update to add hobbies**

db.students.updateMany(

{ "Name": "Eve" },

{ $set: { "hobby": "Dancing" } }

)

```
Atlas atlas-wanmtx-shard-0 [primary] Students> db.students.updateMany(
...     { "Name": "Eve" },
...     { $set: { "hobby": "Dancing" } }
... )
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
```

5. **Find documents where hobbies is set neither to Chess nor to Skating**

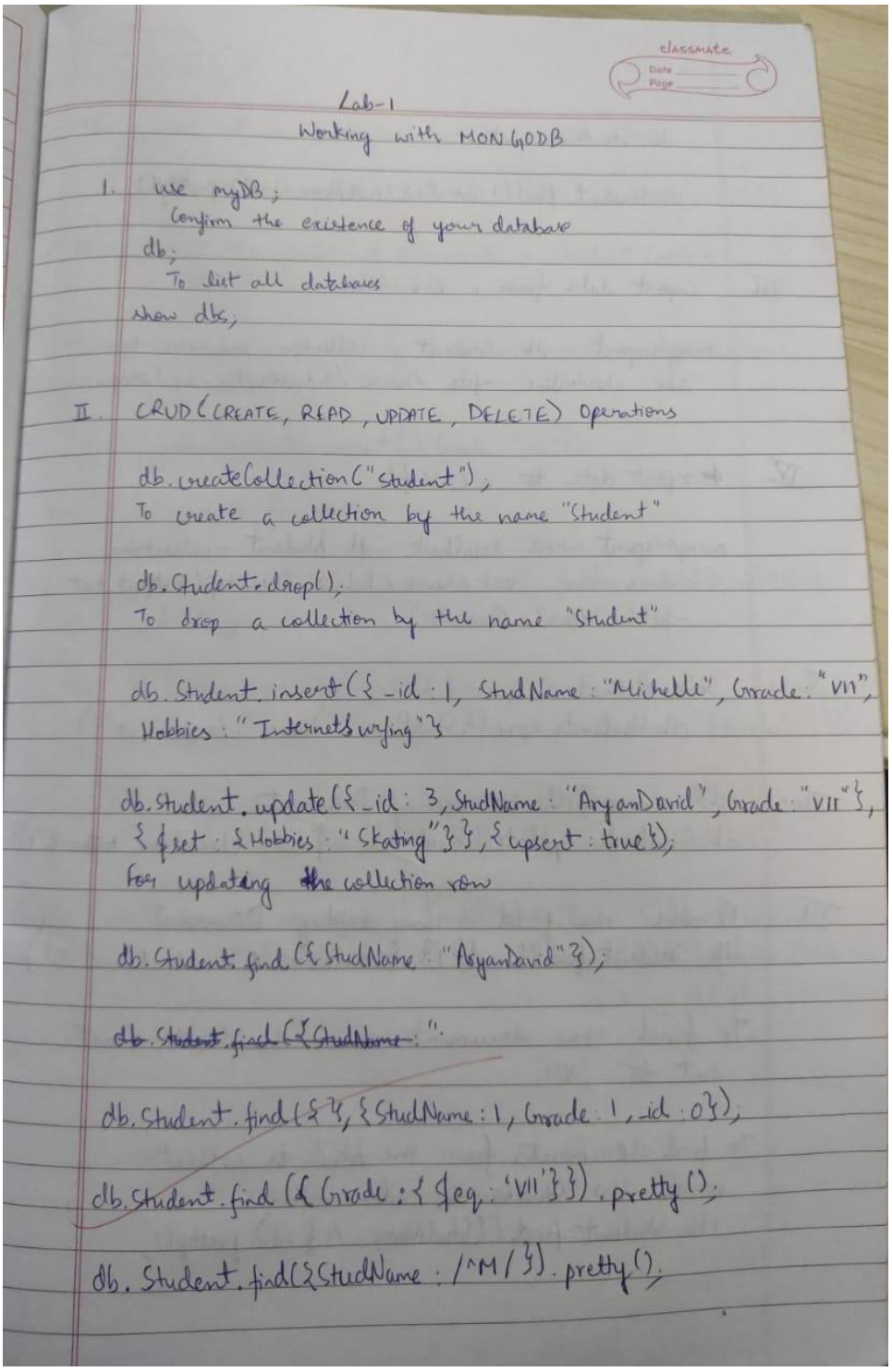
db.students.find({ "hobby": { $nin: ["Chess", "Skating"] } })

```
Atlas atlas-wanmtx-shard-0 [primary] Students> db.students.find({ "hobby": { $nin: ["Chess
", "Skating"] } })
[
  {
    _id: ObjectId("661ce9dc76a00ff8cc51dae1"),
    Rollno: 10,
    Name: 'John',
    Age: 20,
    ContactNo: '1234567890',
    'Email-Id': 'john.doe@example.com',
    grade: 'A',
    hobby: 'Reading'
  },
  {
    _id: ObjectId("661ce9dc76a00ff8cc51dae2"),
    Rollno: 11,
    Name: 'Alicee',
    Age: 21,
    ContactNo: '9876543210',
    'Email-Id': 'alice@example.com',
    grade: 'B',
    hobby: 'Painting'
  },
  {
    _id: ObjectId("661ce9dc76a00ff8cc51dae3"),
    Rollno: 12,
    Name: 'Bob',
    Age: 22,
    ContactNo: '2345678901',
    'Email-Id': 'bob@example.com',
    grade: 'C',
    hobby: 'Cooking'
  },
```

6. Find documents whose name begins with A

db.students.find({ "Name": /^A/ })

```
Atlas atlas-wanmtx-shard-0 [primary] Students> db.students.find({ "Name": /^A/ })
[
  {
    _id: ObjectId("661ce9dc76a00ff8cc51dae2"),
    Rollno: 11,
    Name: 'Alicee',
    Age: 21,
    ContactNo: '9876543210',
    'Email-Id': 'alice@example.com',
    grade: 'B',
    hobby: 'Painting'
  }
]
```



Lab-1

Working with MONGODB

I. use myDB;
Confirm the existence of your database
db;
To list all databases
show dbs;

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

db.createCollection("Student");
To create a collection by the name "Student"

db.Student.drop();
To drop a collection by the name "Student"

db.Student.insert({_id:1, StudName:"Michelle", Grade:"VII", Hobbies:"Internet surfing"})

db.Student.update({_id:3, StudName:"AryanDavid", Grade:"VII"}, {$set:{Hobbies:"Skating"}}, {upsert:true});
For updating the collection row

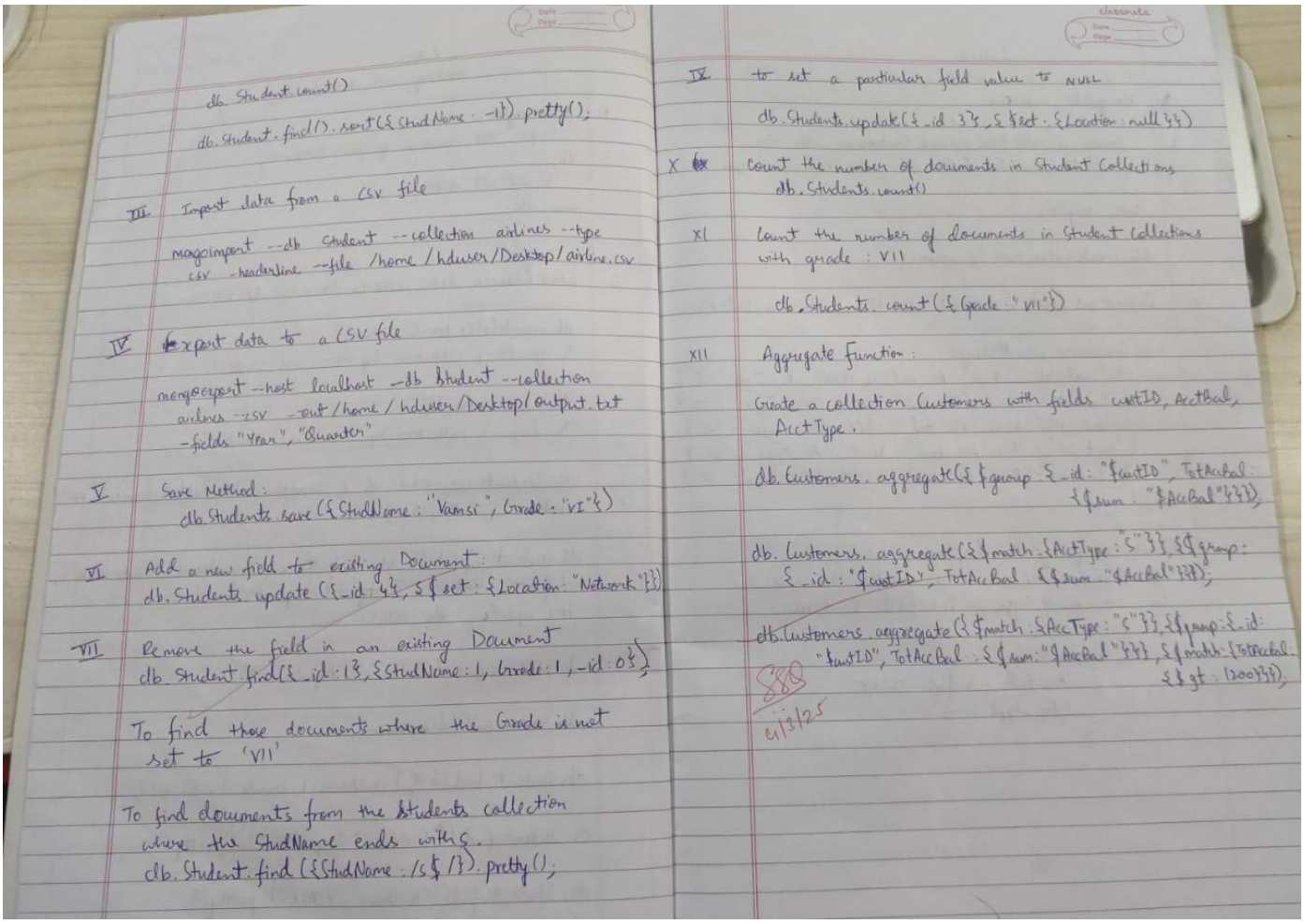
db.Students.find({StudName:"AryanDavid"});

~~db.Student.find({StudName:~~

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

db.Student.find({Grade:{$eq:'VII'}}).pretty();

db.Student.find({StudName:/^M/}).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 --collection airlines --csv --out /home/hduser/Desktop/output.txt --fields "Year", "Quarter"

V Save Method:

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

VI Add a new field to existing Document:

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

VII Remove the field in an existing Document

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

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

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

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

IX to set a particular field value to NULL

db.Students.update({_id:3}, {$set: {Location: null}})

X Count the number of documents in Student Collections

db.Students.count()

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

db.Students.count({Grade: "VII"})

XII Aggregate Function:

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

db.Customers.aggregate({$group: {_id: "$custID", TotAcctBal: {$sum: "$AcctBal"}}});

db.Customers.aggregate({$match: {AcctType: "S"}}, {$group: {_id: "$custID", TotAcctBal: {$sum: "$AcctBal"}}});

db.Customers.aggregate({$match: {AcctType: "S"}}, {$group: {_id: "$custID", TotAcctBal: {$sum: "$AcctBal"}}}, {$match: {TotAcctBal: {$gt: 1200}}});

SSS

4/3/25

Lab 2: 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

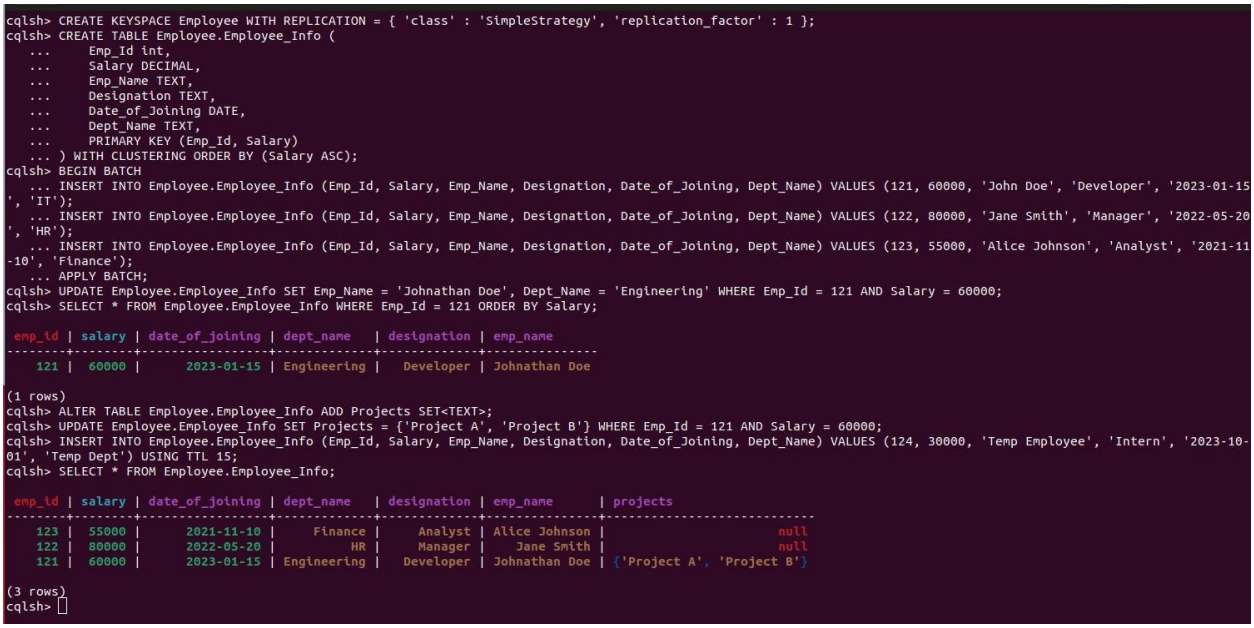
5. Sort the details of Employee records based on salary

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:



```
cqlsh> CREATE KEYSPACE Employee WITH REPLICATION = { 'class' : 'SimpleStrategy', 'replication_factor' : 1 };
cqlsh> CREATE TABLE Employee.Employee_Info (
   ...     Emp_Id int,
   ...     Salary DECIMAL,
   ...     Emp_Name TEXT,
   ...     Designation TEXT,
   ...     Date_of_Joining DATE,
   ...     Dept_Name TEXT,
   ...     PRIMARY KEY (Emp_Id, Salary)
   ... ) WITH CLUSTERING ORDER BY (Salary ASC);
cqlsh> BEGIN BATCH
   ... INSERT INTO Employee.Employee_Info (Emp_Id, Salary, Emp_Name, Designation, Date_of_Joining, Dept_Name) VALUES (121, 60000, 'John Doe', 'Developer', '2023-01-15
', 'IT');
   ... INSERT INTO Employee.Employee_Info (Emp_Id, Salary, Emp_Name, Designation, Date_of_Joining, Dept_Name) VALUES (122, 80000, 'Jane Smith', 'Manager', '2022-05-20
', 'HR');
   ... INSERT INTO Employee.Employee_Info (Emp_Id, Salary, Emp_Name, Designation, Date_of_Joining, Dept_Name) VALUES (123, 55000, 'Alice Johnson', 'Analyst', '2021-11
-10', 'Finance');
   ... APPLY BATCH;
cqlsh> UPDATE Employee.Employee_Info SET Emp_Name = 'Johnathan Doe', Dept_Name = 'Engineering' WHERE Emp_Id = 121 AND Salary = 60000;
cqlsh> SELECT * FROM Employee.Employee_Info WHERE Emp_Id = 121 ORDER BY Salary;

 emp_id | salary | date_of_joining | dept_name   | designation | emp_name
--------+--------+-----------------+-------------+-------------+--------------
    121 |  60000 |      2023-01-15 | Engineering |   Developer | Johnathan Doe

(1 rows)
cqlsh> ALTER TABLE Employee.Employee_Info ADD Projects SET<TEXT>;
cqlsh> UPDATE Employee.Employee_Info SET Projects = {'Project A', 'Project B'} WHERE Emp_Id = 121 AND Salary = 60000;
cqlsh> INSERT INTO Employee.Employee_Info (Emp_Id, Salary, Emp_Name, Designation, Date_of_Joining, Dept_Name) VALUES (124, 30000, 'Temp Employee', 'Intern', '2023-10-
01', 'Temp Dept') USING TTL 15;
cqlsh> SELECT * FROM Employee.Employee_Info;

 emp_id | salary | date_of_joining | dept_name   | designation | emp_name      | projects
--------+--------+-----------------+-------------+-------------+---------------+--------------------------
    123 |  55000 |      2021-11-10 |     Finance |     Analyst | Alice Johnson |                     null
    122 |  80000 |      2022-05-20 |          HR |     Manager |    Jane Smith |                     null
    121 |  60000 |      2023-01-15 | Engineering |   Developer | Johnathan Doe | {'Project A', 'Project B'}

(3 rows)
cqlsh>
```

Lab-2

I MongoDB exercise

1. Create a collection named Customers with the specified attributes (Cust_id, Acc_Bal, Acc_Type):

```
use myDB;

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

2. Insert at Least 5 records into the customers collection:

```
db.Customers.insertMany([
  {"Cust_id": 1, "Acc_Bal": 1500, "Acc_Type": "Z"},
  {"Cust_id": 2, "Acc_Bal": 1200, "Acc_Type": "Z"},
  {"Cust_id": 3, "Acc_Bal": 1300, "Acc_Type": "X"},
  {"Cust_id": 4, "Acc_Bal": 800, "Acc_Type": "Z"},
  {"Cust_id": 5, "Acc_Bal": 2000, "Acc_Type": "Z"}
]);
```

3. Write a query to display records where the total account balance is greater than 1200 for account type z for each Cust_id:

```
db.Customers.find({
    "Acc_Bal": {$gt: 1200},
    "Acc_type": "Z"
});
```

4. Determine Minimum and Maximum account balance for each Cust_id:

```
db.Customers.aggregate([
  {$group: {
      _id: "$Cust_id",
      min_balance: {$min: "$Acc_Bal"},
      max_balance: {$max: "$Acc_Bal"}
    }
  }]);
```

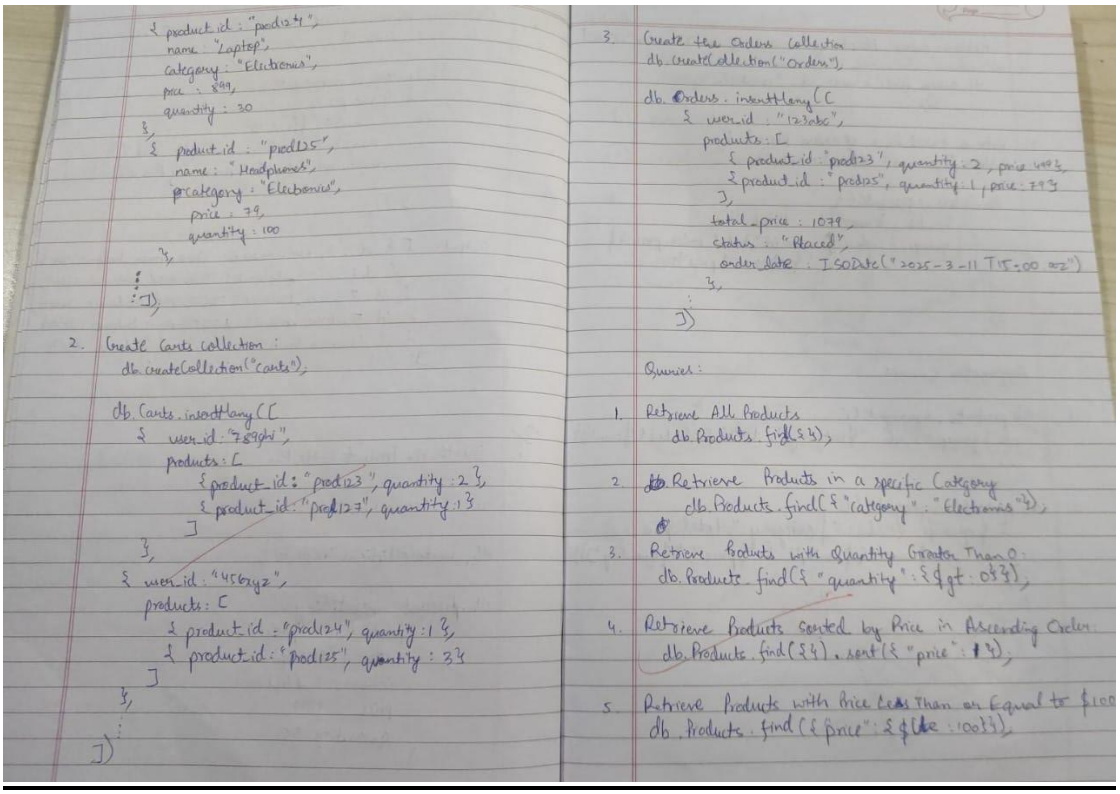
Output:

```
[{_id: 3, min_balance: 1300, max_balance: 1300},
 {_id: 1, min_balance: 1500, max_balance: 1500},
 {_id: 2, min_balance: 1200, max_balance: 1200},
 {_id: 5, min_balance: 2000, max_balance: 2000},
 {_id: 4, min_balance: 800, max_balance: 800}]
```

II

1. Create the Products Collection

```
use ecommerce;

db.createCollection("Products");

db.Products.insertMany([
  {product_id: "prod123",
   name: "Smartphone",
   category: "Electronics",
   price: 499,
   quantity: 50
  },
```



{ product_id : "prod124",
name "Laptop",
category : "Electronics",
price : 899,
quantity : 30
},
{ product_id : "prod125",
name : "Headphones",
category : "Electronics",
price : 79,
quantity : 100
},
:
:])

2. Create Carts collection :
db.createCollection("carts");

db.Carts.insertMany([
{ user_id: "789ghi",
products: [
{ product_id: "prod123", quantity: 2 },
{ product_id: "prod124", quantity: 1 }
]
},
{ user_id: "456xyz",
products: [
{ product_id: "prod124", quantity: 1 },
{ product_id: "prod125", quantity: 3 }
]
},
:
])

3. Create the Orders Collection
db.createCollection("Orders");

db.Orders.insertMany([
{ user_id : "123abc",
products: [
{ product_id "prod123", quantity: 2, price: 499 },
{ product_id : "prod125", quantity: 1, price: 79 }
],
total_price : 1079,
status : "Placed",
order_date : ISODate("2025-3-11T15:00:00Z")
},
:
])

Queries :

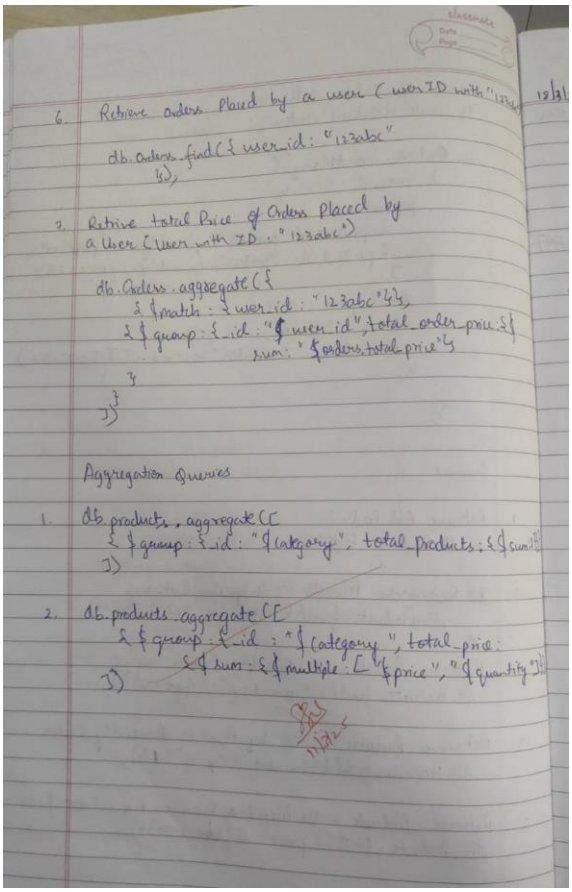
1. Retrieve All Products
db.Products.find({});

2. Retrieve Products in a specific Category
db.Products.find({ "category" : "Electronics" });

3. Retrieve Products with Quantity Greater Than 0:
db.Products.find({ "quantity" : { $gt: 0 } });

4. Retrieve Products sorted by Price in Ascending Order:
db.Products.find({}).sort({ "price" : 1 });

5. Retrieve Products with Price Less Than or Equal to $100
db.Products.find({ "price" : { $lte : 100 } });



6. Retrieve orders Placed by a user (userID with "123abc")

db.Orders.find({ user_id: "123abc" });

7. Retrive total Price of Orders Placed by a User (User with ID : "123abc")

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

Aggregation Queries

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

2. db.products.aggregate([
{ $group : { _id : "$category", total_price:
{ $sum : { $multiply : ["$price", "$quantity"] }
])

Lab 3: 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.

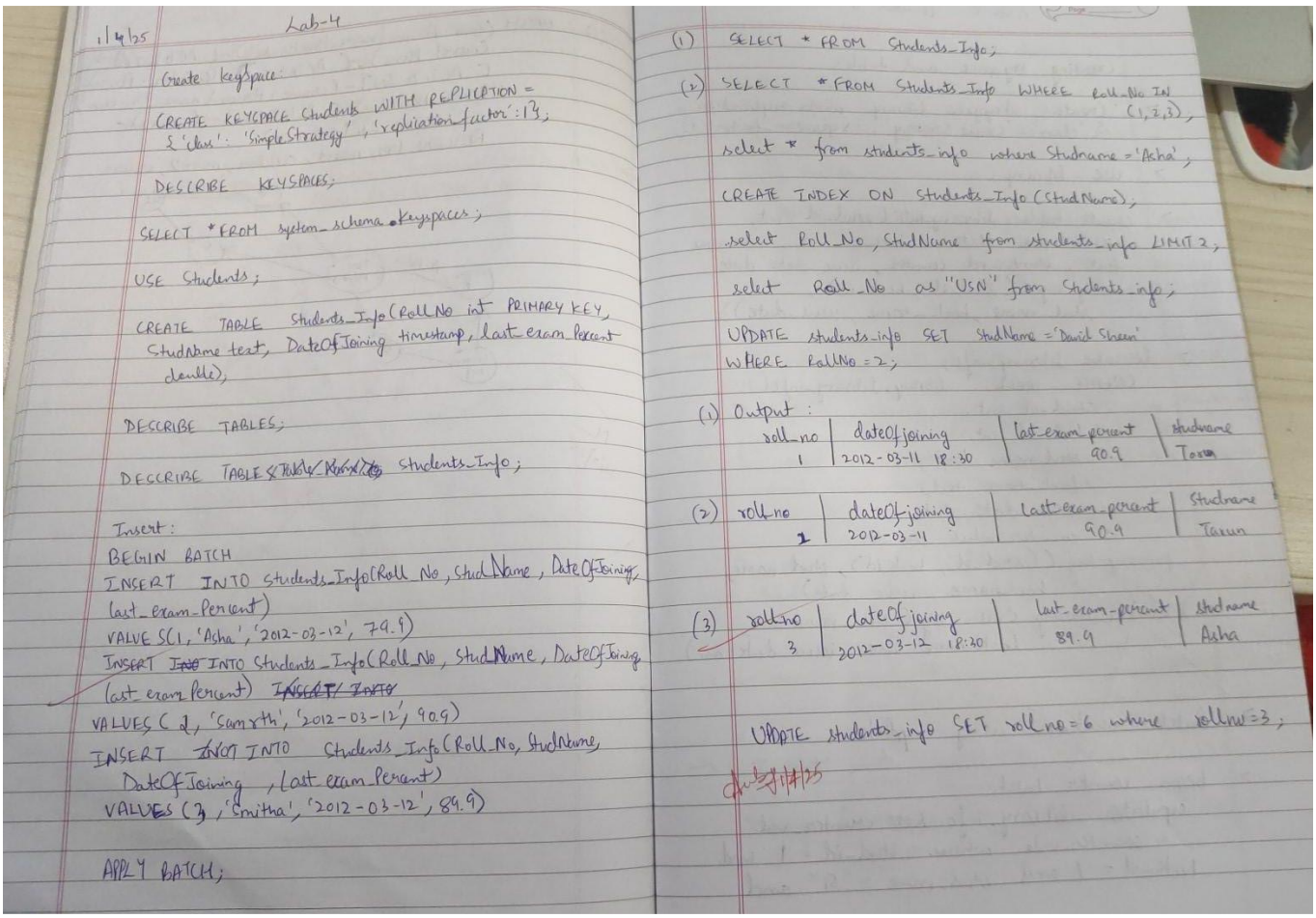
6. Export the created column to a csv file

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

Code with Output:

```
cqlsh> CREATE KEYSPACE Library WITH REPLICATION = { 'class' : 'SimpleStrategy', 'replication_factor' : 1 };
cqlsh> CREATE TABLE Library.Library_Info (
   ...     Stud_Id int,
   ...     Book_Name TEXT,
   ...     Book_Id int,
   ...     Date_of_issue DATE,
   ...     PRIMARY KEY (Stud_Id, Book_Name, Date_of_issue)
   ... );
cqlsh> BEGIN BATCH
   ... INSERT INTO Library.Library_Info (Stud_Id, Book_Name, Book_Id, Date_of_issue) VALUES (112, 'BDA', 1, '2023-09-01');
   ... INSERT INTO Library.Library_Info (Stud_Id, Book_Name, Book_Id, Date_of_issue) VALUES (112, 'BDA', 1, '2023-09-05');
   ... INSERT INTO Library.Library_Info (Stud_Id, Book_Name, Book_Id, Date_of_issue) VALUES (113, 'ML', 2, '2023-09-02');
   ... INSERT INTO Library.Library_Info (Stud_Id, Book_Name, Book_Id, Date_of_issue) VALUES (114, 'AI', 3, '2023-09-03');
   ... INSERT INTO Library.Library_Info (Stud_Id, Book_Name, Book_Id, Date_of_issue) VALUES (115, 'DBMS', 4, '2023-09-04');
   ... APPLY BATCH;
cqlsh> SELECT * FROM Library.Library_Info;

 stud_id | book_name | date_of_issue | book_id
---------+-----------+---------------+---------
     114 |        AI |    2023-09-03 |       3
     113 |        ML |    2023-09-02 |       2
     112 |       BDA |    2023-09-01 |       1
     112 |       BDA |    2023-09-05 |       1
     115 |      DBMS |    2023-09-04 |       4

(5 rows)
cqlsh> SELECT COUNT(*) FROM Library.Library_Info WHERE Stud_Id = 112 AND Book_Name = 'BDA';

 count
-------
     2

(1 rows)
```

```
cqlsh> COPY Library.Library_Info TO 'library_info.csv' WITH HEADER = TRUE;
Using 16 child processes

Starting copy of library.library_info with columns [stud_id, book_name, date_of_issue, book_id].
Processed: 5 rows; Rate:      96 rows/s; Avg. rate:      96 rows/s
5 rows exported to 1 files in 0.089 seconds.
cqlsh> COPY Library.Library_Info FROM 'library_info.csv' WITH HEADER = TRUE;
Using 16 child processes

Starting copy of library.library_info with columns [stud_id, book_name, date_of_issue, book_id].
Processed: 5 rows; Rate:       9 rows/s; Avg. rate:      13 rows/s
5 rows imported from 1 files in 0.375 seconds (0 skipped).
cqlsh>
```



Lab 4: Cassandra

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

Code with Output:

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ cd ./Desktop/
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ 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-800-G9-Desktop-PC]
Starting resourcemanager
Starting nodemanagers
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -mkdir /Lab05
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hadoop fs -ls /Hadoop
ls: `/Hadoop': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hadoop fs -ls /Lab05
```

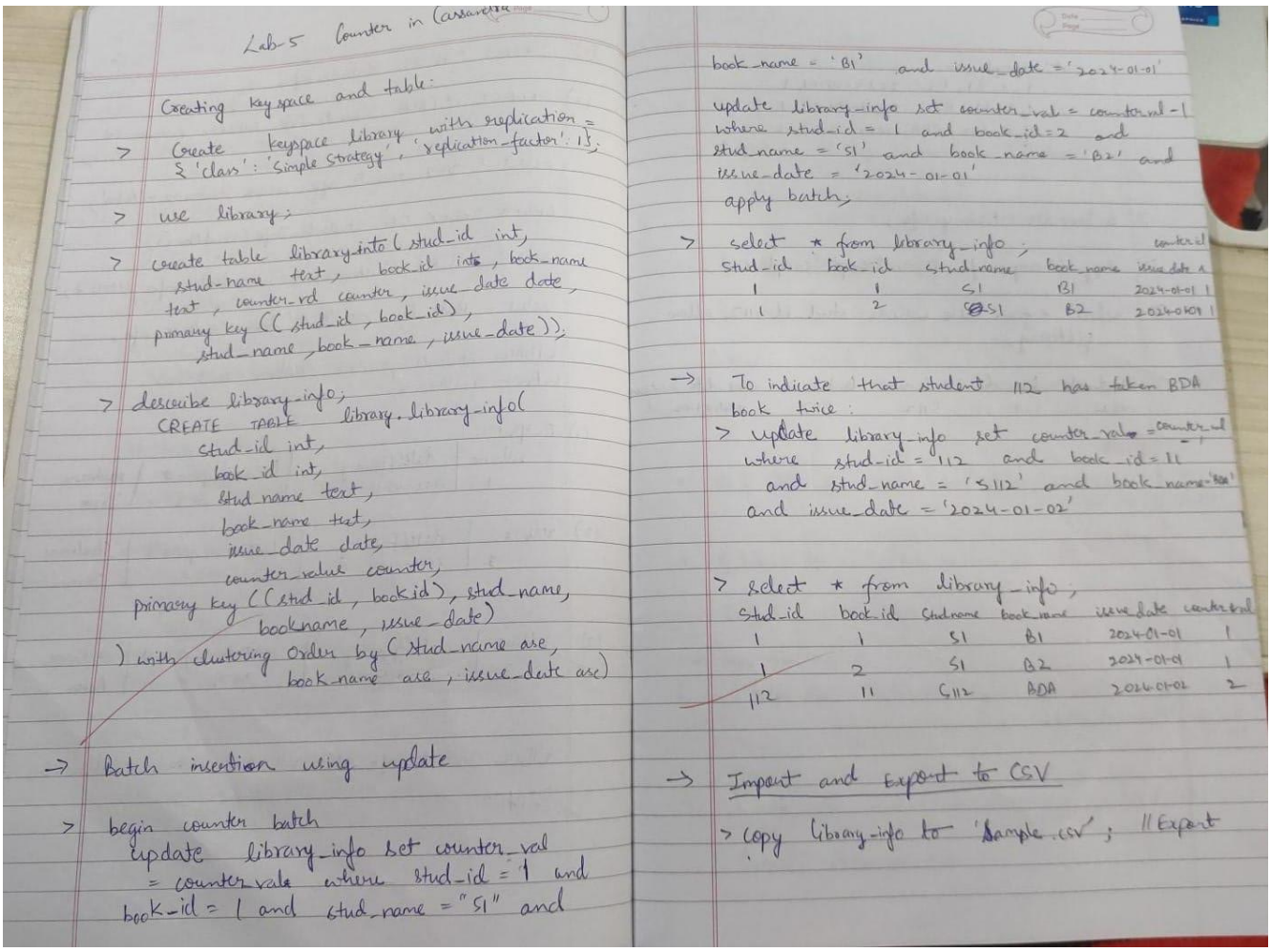
```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ touch test.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ nano text.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -put ./text.txt /Lab05/text.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hadoop fs -ls /Lab05
Found 1 items
-rw-r--r--   1 hadoop supergroup         19 2024-05-13 14:33 /Lab05/text.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -cat /Lab05/text.txt
Hello
How are you?
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hadoop fs -ls /Lab05
Found 2 items
-rw-r--r--   1 hadoop supergroup         15 2024-05-13 14:40 /Lab05/test.txt
-rw-r--r--   1 hadoop supergroup         19 2024-05-13 14:33 /Lab05/text.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -getmerge /Lab05 /text.txt /Lab05 /test.txt ../
Downloads/Merged.txt
getmerge: `/text.txt': No such file or directory
getmerge: `/test.txt': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -getmerge /Lab05/text.txt /Lab05/test.txt ../Do
wnloads/Merged.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hadoop fs -getfacl /Lab05
# file: /Lab05
# owner: hadoop
# group: supergroup
user::rwx
group::r-x
other::r-x
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -copyToLocal /Lab05/text.txt ../Documents
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -copyToLocal /Lab05/test.txt ../Documents
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -cat /Lab05/text.txt
Hello
How are you?
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -mv /Lab05 /test_Lab05
```

```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -ls /test_Lab05
Found 2 items
-rw-r--r--   1 hadoop supergroup         15 2024-05-13 14:40 /test_Lab05/test.txt
-rw-r--r--   1 hadoop supergroup         19 2024-05-13 14:33 /test_Lab05/text.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -cp /test_Lab05/ /Lab05
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -ls /Lab05
Found 2 items
-rw-r--r--   1 hadoop supergroup         15 2024-05-13 14:51 /Lab05/test.txt
-rw-r--r--   1 hadoop supergroup         19 2024-05-13 14:51 /Lab05/text.txt
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -ls /test_Lab05
Found 2 items
-rw-r--r--   1 hadoop supergroup         15 2024-05-13 14:40 /test_Lab05/test.txt
-rw-r--r--   1 hadoop supergroup         19 2024-05-13 14:33 /test_Lab05/text.txt
```



Lab-5 Counter in Cassandra

Creating keyspace and table:

> Create keyspace library with replication = {'class': 'SimpleStrategy', 'replication_factor': 1};

> use library;

> create table library_info (stud_id int, stud_name text, book_id int, book_name text, counter_val counter, issue_date date, primary key ((stud_id, book_id), stud_name, book_name, issue_date));

> describe library_info;

CREATE TABLE library.library_info(
stud_id int,
book_id int,
stud_name text,
book_name text,
issue_date date,
counter_value counter,
primary key ((stud_id, book_id), stud_name, book_name, issue_date)
) with clustering order by (stud_name asc, book_name asc, issue_date asc)

→ Batch insertion using update

> begin counter batch
update library_info set counter_val = counter_val where stud_id = 1 and book_id = 1 and stud_name = "S1" and book_name = 'B1' and issue_date = '2024-01-01'
update library_info set counter_val = counter_val - 1 where stud_id = 1 and book_id = 2 and stud_name = 'S1' and book_name = 'B2' and issue_date = '2024-01-01'
apply batch;

> select * from library_info;

stud_id	book_id	stud_name	book_name	issue_date	counter_val
1	1	S1	B1	2024-01-01	1
1	2	S1	B2	2024-01-01	1

→ To indicate that student 112 has taken BDA book twice:

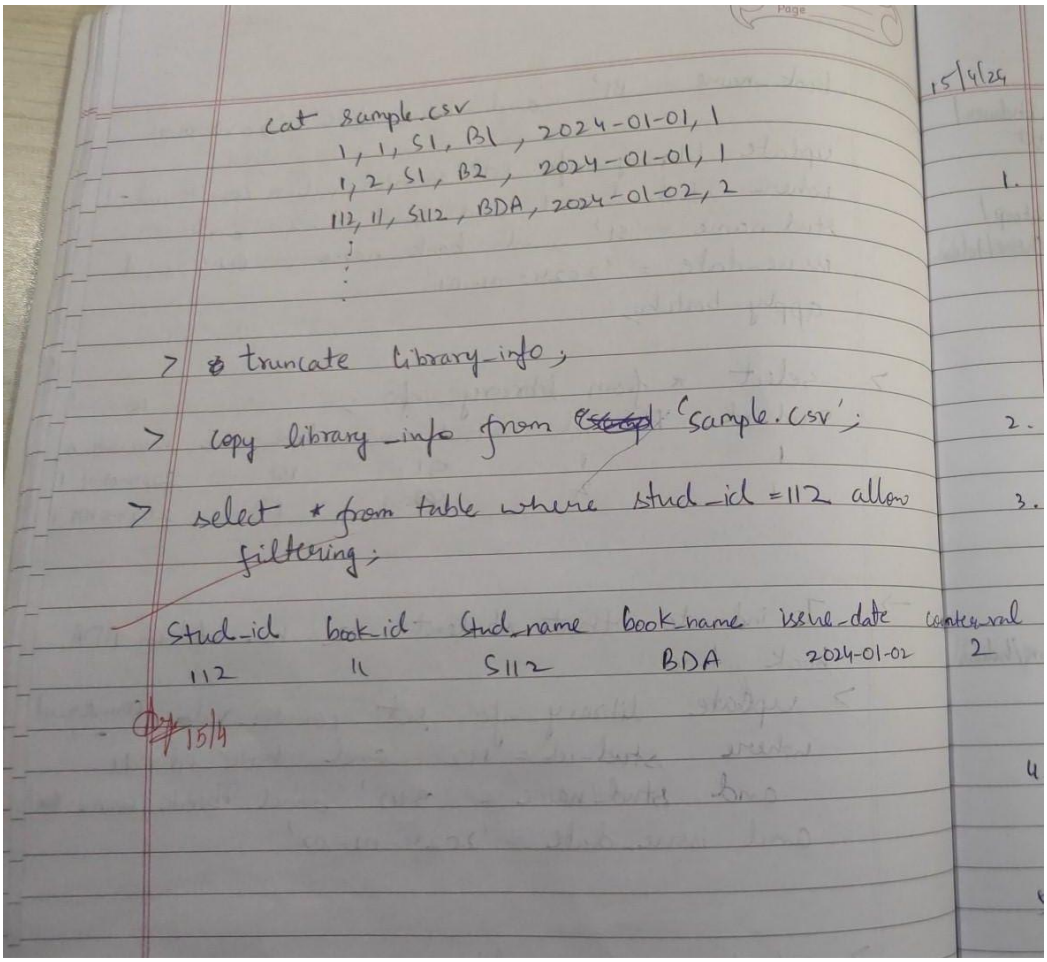
> update library_info set counter_val = counter_val + 1 where stud_id = 112 and book_id = 11 and stud_name = 'S112' and book_name = 'BDA' and issue_date = '2024-01-02'

> select * from library_info;

stud_id	book_id	stud_name	book_name	issue_date	counter_val
1	1	S1	B1	2024-01-01	1
1	2	S1	B2	2024-01-01	1
112	11	S112	BDA	2024-01-02	2

→ Import and export to CSV

> copy library_info to 'sample.csv'; // Export



cat sample.csv

1, 1, S1, B1, 2024-01-01, 1

1, 2, S1, B2, 2024-01-01, 1

112, 11, S112, BDA, 2024-01-02, 2

⋮

15/4/24

1.

> truncate library_info;

2.

> copy library_info from 'sample.csv';

3.

> select * from table where stud_id = 112 allow filtering;

Stud_id	book_id	Stud_name	book_name	issue_date	counter_val
112	11	S112	BDA	2024-01-02	2

15/4

4

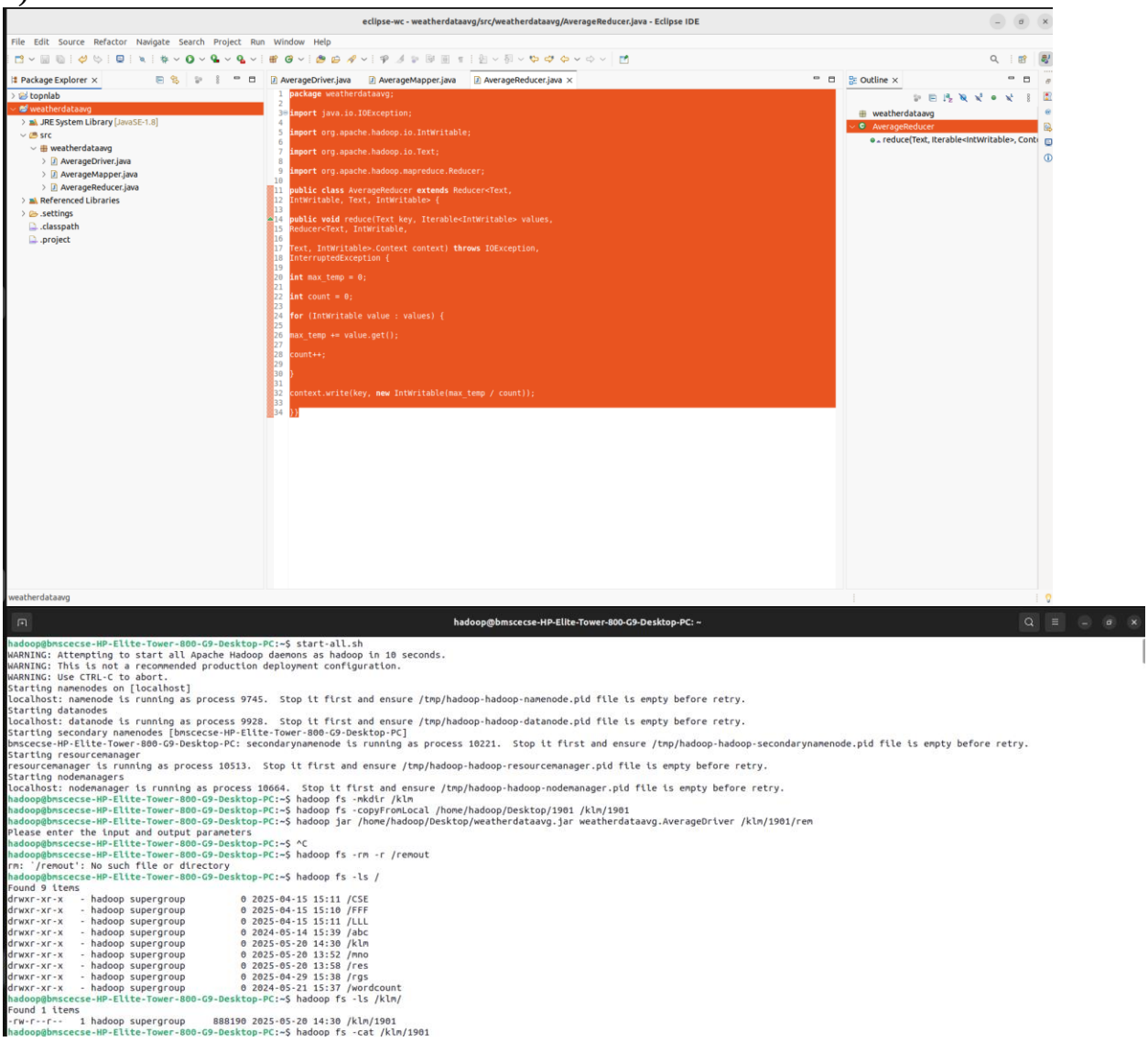
Lab 5: Hadoop

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

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

Code with Output:

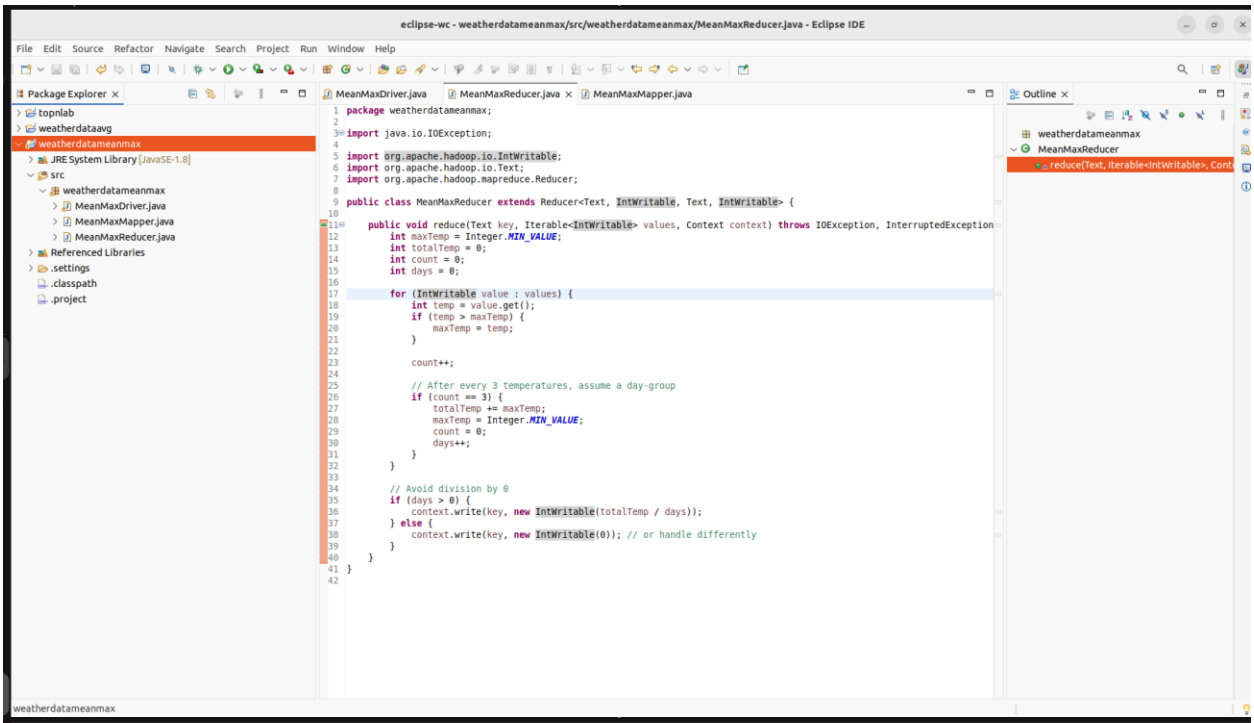
a)

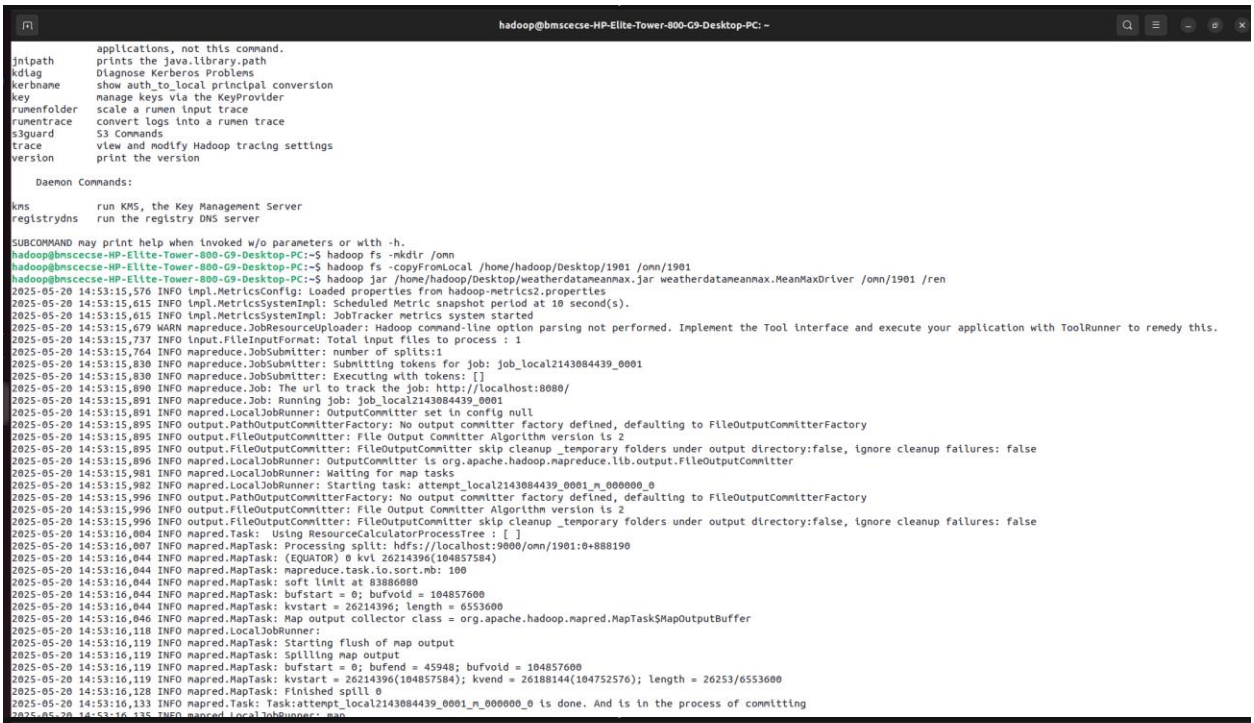


```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/weatherdataavg.jar weatherdataavg.AverageDriver /kln/1901 /rem
2025-05-20 14:34:21,074 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-05-20 14:34:21,116 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-05-20 14:34:21,116 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-05-20 14:34:21,176 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-05-20 14:34:21,230 INFO input.FileInputFormat: Total input files to process : 1
2025-05-20 14:34:21,262 INFO mapreduce.JobSubmitter: number of splits:1
2025-05-20 14:34:21,334 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local261815800_0001
2025-05-20 14:34:21,334 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-20 14:34:21,400 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-05-20 14:34:21,401 INFO mapreduce.Job: Running job: job_local261815800_0001
2025-05-20 14:34:21,402 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-05-20 14:34:21,405 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:34:21,405 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:34:21,406 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:34:21,406 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2025-05-20 14:34:21,446 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-05-20 14:34:21,447 INFO mapred.LocalJobRunner: Starting task: attempt_local261815800_0001_m_000000_0
2025-05-20 14:34:21,459 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:34:21,460 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:34:21,460 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:34:21,469 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-05-20 14:34:21,472 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/kln/1901:0+888190
2025-05-20 14:34:21,511 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-05-20 14:34:21,511 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-05-20 14:34:21,511 INFO mapred.MapTask: soft limit at 83886080
2025-05-20 14:34:21,511 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-05-20 14:34:21,511 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-05-20 14:34:21,513 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-20 14:34:21,586 INFO mapred.LocalJobRunner:
2025-05-20 14:34:21,587 INFO mapred.MapTask: Starting flush of map output
2025-05-20 14:34:21,587 INFO mapred.MapTask: Spilling map output
2025-05-20 14:34:21,587 INFO mapred.MapTask: bufstart = 0; bufend = 59076; bufvoid = 104857600
2025-05-20 14:34:21,587 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26188144(104752576); length = 26253/6553600
2025-05-20 14:34:21,595 INFO mapred.MapTask: Finished spill 0
2025-05-20 14:34:21,600 INFO mapred.Task: Task:attempt_local261815800_0001_m_000000_0 is done. And is in the process of committing
2025-05-20 14:34:21,602 INFO mapred.LocalJobRunner: map
2025-05-20 14:34:21,602 INFO mapred.Task: Task 'attempt_local261815800_0001_m_000000_0' done.
2025-05-20 14:34:21,605 INFO mapred.Task: Final Counters for attempt_local261815800_0001_m_000000_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=4430
                FILE: Number of bytes written=713998
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=888190
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=1
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=6565
                Map output records=6564
                HDFS: Number of bytes written=8
                HDFS: Number of read operations=10
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=72210
                Reduce input records=6564
                Reduce output records=1
                Spilled Records=6564
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=633339904
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=8
2025-05-20 14:34:21,803 INFO mapred.LocalJobRunner: Finishing task: attempt_local261815800_0001_r_000000_0
2025-05-20 14:34:21,803 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-05-20 14:34:22,404 INFO mapreduce.Job: Job job_local261815800_0001 running in uber mode : false
2025-05-20 14:34:22,405 INFO mapreduce.Job:  map 100% reduce 100%
2025-05-20 14:34:22,407 INFO mapreduce.Job: Job job_local261815800_0001 completed successfully
2025-05-20 14:34:22,414 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=153312
                FILE: Number of bytes written=1500206
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1776388
                HDFS: Number of bytes written=8
                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=6565
                Map output records=6564
                Map output bytes=59076
                Map output materialized bytes=72210
                Input split bytes=95
                Combine input records=0
                Combine output records=0
                Reduce input groups=1
                Reduce shuffle bytes=72210
                Reduce input records=6564
```

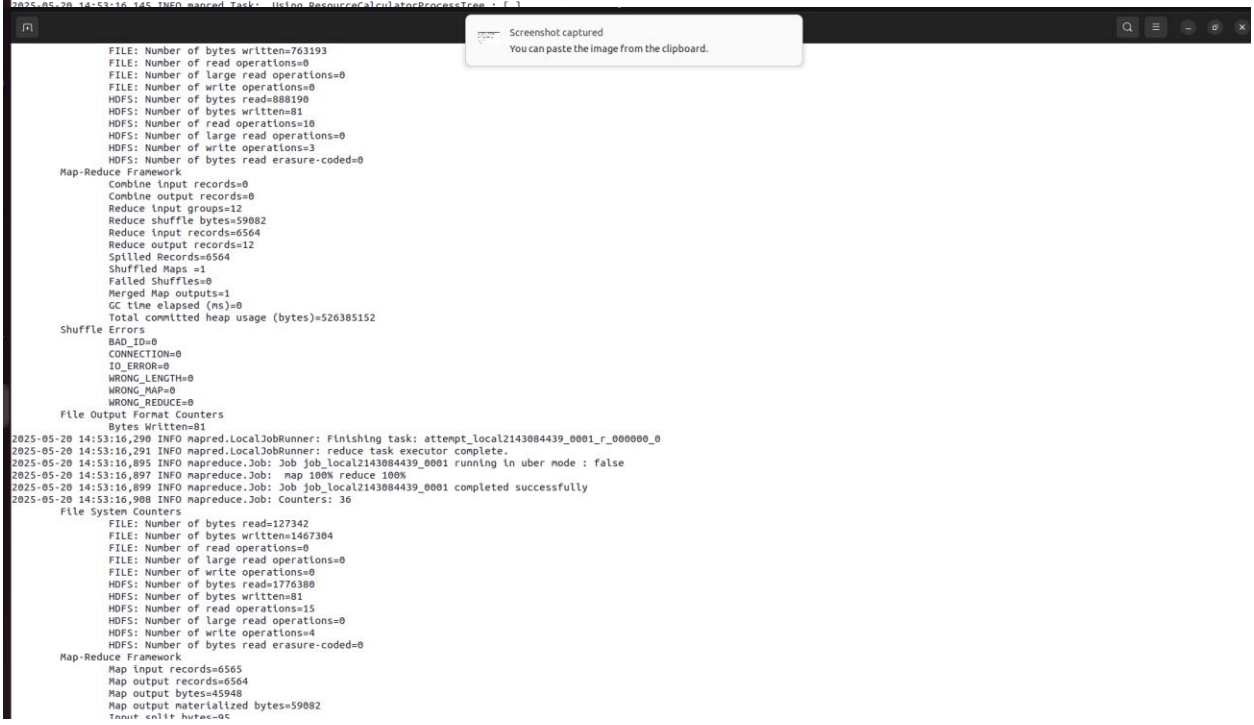
```
                Map input records=6565
                Map output records=6564
                Map output bytes=59076
                Map output materialized bytes=72210
                Input split bytes=95
                Combine input records=0
                Spilled Records=6564
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=633339904
        File Input Format Counters
                Bytes Read=888190
2025-05-20 14:34:21,605 INFO mapred.LocalJobRunner: Finishing task: attempt_local261815800_0001_m_000000_0
2025-05-20 14:34:21,606 INFO mapred.LocalJobRunner: map task executor complete.
2025-05-20 14:34:21,607 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2025-05-20 14:34:21,608 INFO mapred.LocalJobRunner: Starting task: attempt_local261815800_0001_r_000000_0
2025-05-20 14:34:21,612 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:34:21,612 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:34:21,612 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:34:21,612 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-05-20 14:34:21,613 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@c4d4167
2025-05-20 14:34:21,614 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-05-20 14:34:21,622 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=582945312, maxSingleShuffleLimit=145736328, mergeThreshold=3847439360, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-05-20 14:34:21,623 INFO reduce.EventFetcher: attempt_local261815800_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2025-05-20 14:34:21,635 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local261815800_0001_m_000000_0 decomp: 72206 len: 72210 to MEMORY
2025-05-20 14:34:21,637 INFO reduce.InMemoryMapOutput: Read 72206 bytes from map-output for attempt_local261815800_0001_m_000000_0
2025-05-20 14:34:21,637 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 72206, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->72206
2025-05-20 14:34:21,638 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2025-05-20 14:34:21,638 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 14:34:21,638 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
2025-05-20 14:34:21,641 INFO mapred.Merger: Merging 1 sorted segments
2025-05-20 14:34:21,641 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 72199 bytes
2025-05-20 14:34:21,644 INFO reduce.MergeManagerImpl: Merged 1 segments, 72206 bytes to disk to satisfy reduce memory limit
2025-05-20 14:34:21,645 INFO reduce.MergeManagerImpl: Merging 1 files, 72210 bytes from disk
2025-05-20 14:34:21,645 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2025-05-20 14:34:21,645 INFO mapred.Merger: Merging 1 sorted segments
2025-05-20 14:34:21,645 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 72199 bytes
2025-05-20 14:34:21,646 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 14:34:21,679 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2025-05-20 14:34:21,767 INFO mapred.Task: Task:attempt_local261815800_0001_r_000000_0 is done. And is in the process of committing
2025-05-20 14:34:21,771 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 14:34:21,771 INFO mapred.Task: Task attempt_local261815800_0001_r_000000_0 is allowed to commit now
2025-05-20 14:34:21,800 INFO output.FileOutputCommitter: Saved output of task 'attempt_local261815800_0001_r_000000_0' to hdfs://localhost:9000/rem
2025-05-20 14:34:21,802 INFO mapred.LocalJobRunner: reduce > reduce
2025-05-20 14:34:21,802 INFO mapred.Task: Task 'attempt_local261815800_0001_r_000000_0' done.
2025-05-20 14:34:21,803 INFO mapred.Task: Final Counters for attempt_local261815800_0001_r_000000_0: Counters: 30
        File System Counters
                FILE: Number of bytes read=148882
                FILE: Number of bytes written=786208
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=888190
                HDFS: Number of bytes written=8
                HDFS: Number of read operations=10
                Reduce shuffle bytes=72210
                Reduce input records=6564
                Reduce output records=1
                Spilled Records=13128
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=1266679808
        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-800-G9-Desktop-PC:~$ hadoop fs -cat /rem/part-00000
cat: `/rem/part-00000': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /rem/part-r-00000
1901    46
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ package weatherdataavg;

import java.io.IOException;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Reducer;

public class AverageReducer extends Reducer<Text,
IntWritable, Text, IntWritable> {

public void reduce(Text key, Iterable<IntWritable> values,
Reducer<Text, IntWritable,

Text, IntWritable>.Context context) throws IOException,
InterruptedException {

int max_temp = 0;

int count = 0;

for (IntWritable value : values) {

max_temp += value.get();

count++;

}

context.write(key, new IntWritable(max_temp / count));
```

b)


eclipse-wc - weatherdatameanmax/src/weatherdatameanmax/MeanMaxReducer.java - Eclipse IDE
File Edit Source Refactor Navigate Search Project Run Window Help
Package Explorer
topnlab
weatherdataavg
weatherdatameanmax
JRE System Library [JavaSE-1.8]
src
weatherdatameanmax
MeanMaxDriver.java
MeanMaxMapper.java
MeanMaxReducer.java
Referenced Libraries
.settings
.classpath
.project
MeanMaxDriver.java
MeanMaxReducer.java
MeanMaxMapper.java
package weatherdatameanmax;
import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
public class MeanMaxReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
public void reduce(Text key, Iterable<IntWritable> values, Context context) throws IOException, InterruptedException
int maxTemp = Integer.MIN_VALUE;
int totalTemp = 0;
int count = 0;
int days = 0;
for (IntWritable value : values) {
int temp = value.get();
if (temp > maxTemp) {
maxTemp = temp;
}
count++;
// After every 3 temperatures, assume a day-group
if (count == 3) {
totalTemp += maxTemp;
maxTemp = Integer.MIN_VALUE;
count = 0;
days++;
}
}
// Avoid division by 0
if (days > 0) {
context.write(key, new IntWritable(totalTemp / days));
} else {
context.write(key, new IntWritable(0)); // or handle differently
}
}
}
Outline
weatherdatameanmax
MeanMaxReducer
reduce(Text, Iterable<IntWritable>, Cont
weatherdatameanmax
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
hadoop@bmscecse-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]
localhost: namenode is running as process 9745. Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 9928. Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondary namenodes [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: secondarynamenode is running as process 10221. Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
Starting resourcemanager
resourcemanager is running as process 10513. Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 10664. 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 ls -mkdir /omn
ERROR: ls is not COMMAND nor fully qualified CLASSNAME.
Usage: hadoop [OPTIONS] SUBCOMMAND [SUBCOMMAND OPTIONS]
or hadoop [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
where CLASSNAME is a user-provided Java class
OPTIONS is none or any of:
buildpaths attempt to add class files from build tree
--config dir Hadoop config directory
--debug turn on shell script debug mode
--help usage information
hostnames list[,of,host,names] hosts to use in worker mode
hosts filename list of hosts to use in worker mode
loglevel level set the log4j level for this command
workers turn on worker mode
SUBCOMMAND is one of:
Admin Commands:
daemonlog get/set the log level for each daemon
Client Commands:
archive create a Hadoop archive
checknative check native Hadoop and compression libraries availability
classpath prints the class path needed to get the Hadoop jar and the
required libraries
conftest validate configuration XML files
credential interact with credential providers
distch distributed metadata changer
distcp copy file or directories recursively
dtutil operations related to delegation tokens
envvars display computed Hadoop environment variables
fs run a generic filesystem user client
gridmix submit a mix of synthetic job, modeling a profiled from
production load
jar <jar> run a jar file. NOTE: please use "yarn jar" to launch YARN
applications, not this command.

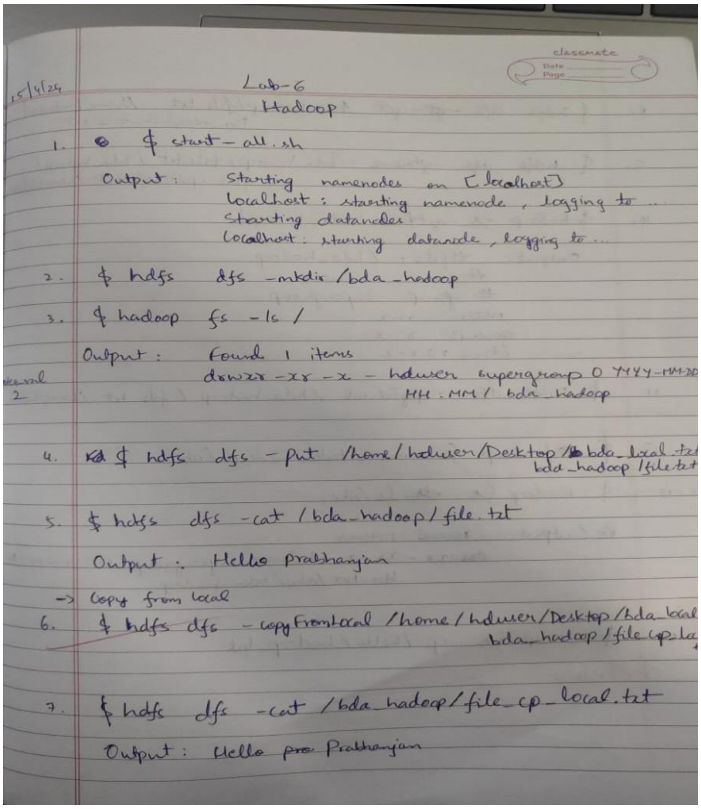

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
applications, not this command.
jnipath prints the java.library.path
kdiag Diagnose Kerberos Problems
kerbname show auth_to_local principal conversion
key manage keys via the KeyProvider
rumenfolder scale a rumen input trace
rumentrace convert logs into a rumen trace
s3guard S3 Commands
trace view and modify Hadoop tracing settings
version print the version
Daemon Commands:
kms run KMS, the Key Management Server
registrydns run the registry DNS server
SUBCOMMAND may print help when invoked w/o parameters or with -h.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mkdir /omn
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -copyFromLocal /home/hadoop/Desktop/1901 /omn/1901
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/weatherdatameanmax.jar weatherdatameanmax.MeanMaxDriver /omn/1901 /ren
2025-05-20 14:53:15,576 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-05-20 14:53:15,615 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-05-20 14:53:15,615 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-05-20 14:53:15,679 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2025-05-20 14:53:15,737 INFO input.FileInputFormat: Total input files to process : 1
2025-05-20 14:53:15,764 INFO mapreduce.JobSubmitter: number of splits:1
2025-05-20 14:53:15,830 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2143084439_0001
2025-05-20 14:53:15,830 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-20 14:53:15,890 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-05-20 14:53:15,891 INFO mapreduce.Job: Running job: job_local2143084439_0001
2025-05-20 14:53:15,891 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-05-20 14:53:15,895 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:53:15,895 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:53:15,895 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:53:15,896 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2025-05-20 14:53:15,981 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-05-20 14:53:15,982 INFO mapred.LocalJobRunner: Starting task: attempt_local2143084439_0001_m_000000_0
2025-05-20 14:53:15,996 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:53:15,996 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:53:15,996 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:53:16,004 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2025-05-20 14:53:16,007 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/omn/1901:0+888190
2025-05-20 14:53:16,044 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-05-20 14:53:16,044 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-05-20 14:53:16,044 INFO mapred.MapTask: soft limit at 83886080
2025-05-20 14:53:16,044 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-05-20 14:53:16,044 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-05-20 14:53:16,046 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-20 14:53:16,118 INFO mapred.LocalJobRunner:
2025-05-20 14:53:16,119 INFO mapred.MapTask: Starting flush of map output
2025-05-20 14:53:16,119 INFO mapred.MapTask: Spilling map output
2025-05-20 14:53:16,119 INFO mapred.MapTask: bufstart = 0; bufend = 45948; bufvoid = 104857600
2025-05-20 14:53:16,119 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26188144(104752576); length = 26253/6553600
2025-05-20 14:53:16,128 INFO mapred.MapTask: Finished spill 0
2025-05-20 14:53:16,133 INFO mapred.Task: Task:attempt_local2143084439_0001_m_000000_0 is done. And is in the process of committing
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
2025-05-20 14:53:16,145 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:53:16,145 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2025-05-20 14:53:16,146 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@6b64ffe0
2025-05-20 14:53:16,147 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-05-20 14:53:16,155 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=5829453312, maxSingleShuffleLimit=1457363328, mergeThreshold=3847439360, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-05-20 14:53:16,156 INFO reduce.EventFetcher: attempt_local2143084439_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2025-05-20 14:53:16,170 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local2143084439_0001_m_000000_0 decomp: 59078 len: 59082 to MEMORY
2025-05-20 14:53:16,172 INFO reduce.InMemoryMapOutput: Read 59078 bytes from map-output for attempt_local2143084439_0001_m_000000_0
2025-05-20 14:53:16,172 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 59078, inMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory ->59078
2025-05-20 14:53:16,173 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2025-05-20 14:53:16,174 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 14:53:16,174 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
2025-05-20 14:53:16,176 INFO mapred.Merger: Merging 1 sorted segments
2025-05-20 14:53:16,176 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 59073 bytes
2025-05-20 14:53:16,180 INFO reduce.MergeManagerImpl: Merged 1 segments, 59078 bytes to disk to satisfy reduce memory limit
2025-05-20 14:53:16,181 INFO reduce.MergeManagerImpl: Merging 1 files, 59082 bytes from disk
2025-05-20 14:53:16,181 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2025-05-20 14:53:16,181 INFO mapred.Merger: Merging 1 sorted segments
2025-05-20 14:53:16,182 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 59073 bytes
2025-05-20 14:53:16,182 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 14:53:16,212 INFO Configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
2025-05-20 14:53:16,273 INFO mapred.Task: Task:attempt_local2143084439_0001_r_000000_0 is done. And is in the process of committing
2025-05-20 14:53:16,275 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 14:53:16,275 INFO mapred.Task: Task attempt_local2143084439_0001_r_000000_0 is allowed to commit now
2025-05-20 14:53:16,289 INFO output.FileOutputCommitter: Saved output of task 'attempt_local2143084439_0001_r_000000_0' to hdfs://localhost:9000/ren
2025-05-20 14:53:16,290 INFO mapred.LocalJobRunner: reduce > reduce
2025-05-20 14:53:16,290 INFO mapred.Task: Task 'attempt_local2143084439_0001_r_000000_0' done.
2025-05-20 14:53:16,290 INFO mapred.Task: Final Counters for attempt_local2143084439_0001_r_000000_0: Counters: 30
File System Counters
FILE: Number of bytes read=122769
FILE: Number of bytes written=763193
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=888190
HDFS: Number of bytes written=81
HDFS: Number of read operations=10
HDFS: Number of large read operations=0
HDFS: Number of write operations=3
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Combine input records=0
Combine output records=0
Reduce input groups=12
Reduce shuffle bytes=59082
Reduce input records=6564
Reduce output records=12
Spilled Records=6564
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=526385152
Shuffle Errors
BAD_ID=0
CONNECTION=0

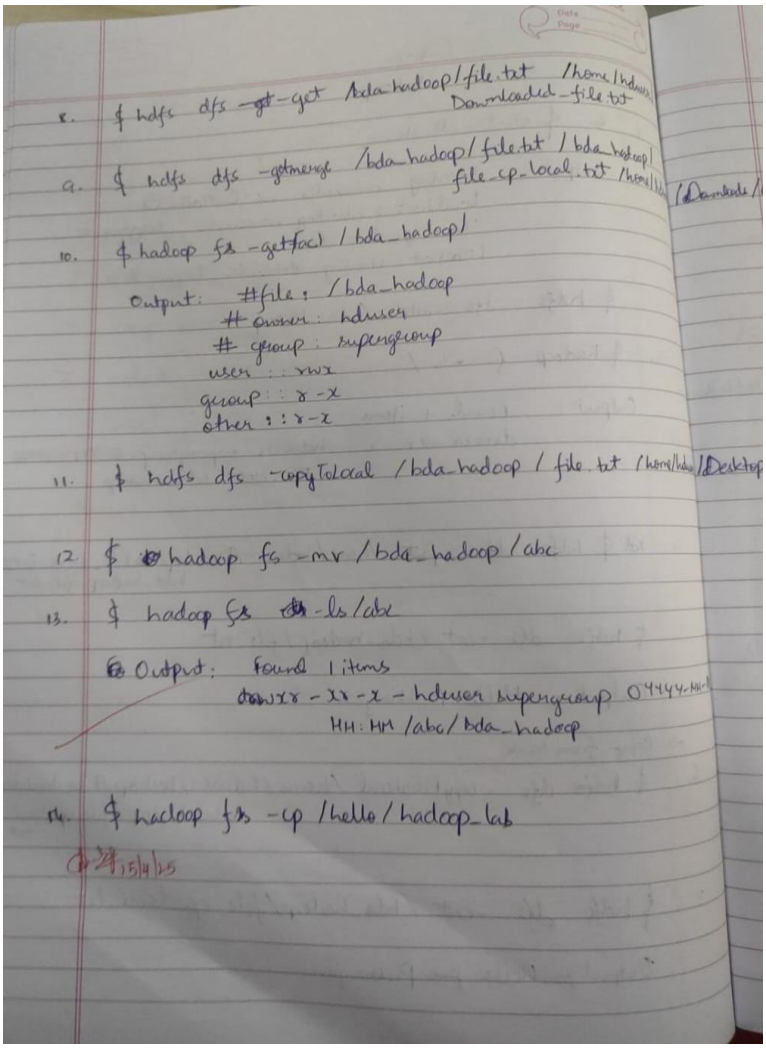
```
Screenshot captured
You can paste the image from the clipboard.
2025-05-20 14:53:15,982 INFO mapred.LocalJobRunner: Starting task: attempt_
2025-05-20 14:53:15,996 INFO output.PathOutputCommitterFactory: No output c                                                                    cterFactory
2025-05-20 14:53:15,996 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:53:15,996 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:53:16,004 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2025-05-20 14:53:16,007 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/omn/1901:0+888190
2025-05-20 14:53:16,044 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-05-20 14:53:16,044 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-05-20 14:53:16,044 INFO mapred.MapTask: soft limit at 83886080
2025-05-20 14:53:16,044 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-05-20 14:53:16,044 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-05-20 14:53:16,046 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-20 14:53:16,118 INFO mapred.LocalJobRunner:
2025-05-20 14:53:16,119 INFO mapred.MapTask: Starting flush of map output
2025-05-20 14:53:16,119 INFO mapred.MapTask: Spilling map output
2025-05-20 14:53:16,119 INFO mapred.MapTask: bufstart = 0; bufend = 45948; bufvoid = 104857600
2025-05-20 14:53:16,119 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26188144(104752576); length = 26253/6553600
2025-05-20 14:53:16,128 INFO mapred.MapTask: Finished spill 0
2025-05-20 14:53:16,133 INFO mapred.Task: Task:attempt_local2143084439_0001_m_000000_0 is done. And is in the process of committing
2025-05-20 14:53:16,135 INFO mapred.LocalJobRunner: map
2025-05-20 14:53:16,135 INFO mapred.Task: Task 'attempt_local2143084439_0001_m_000000_0' done.
2025-05-20 14:53:16,138 INFO mapred.Task: Final Counters for attempt_local2143084439_0001_m_000000_0: Counters: 23
        File System Counters
                FILE: Number of bytes read=4573
                FILE: Number of bytes written=704111
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=888190
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=1
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=6565
                Map output records=6564
                Map output bytes=45948
                Map output materialized bytes=59082
                Input split bytes=95
                Combine input records=0
                Spilled Records=6564
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=526385152
        File Input Format Counters
                Bytes Read=888190
2025-05-20 14:53:16,138 INFO mapred.LocalJobRunner: Finishing task: attempt_local2143084439_0001_m_000000_0
2025-05-20 14:53:16,139 INFO mapred.LocalJobRunner: map task executor complete.
2025-05-20 14:53:16,140 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2025-05-20 14:53:16,140 INFO mapred.LocalJobRunner: Starting task: attempt_local2143084439_0001_r_000000_0
2025-05-20 14:53:16,145 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:53:16,145 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 14:53:16,145 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:53:16,145 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
```



```
Screenshot captured
You can paste the image from the clipboard.
                FILE: Number of bytes written=763193
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=888190
                HDFS: Number of bytes written=81
                HDFS: Number of read operations=10
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=3
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Combine input records=0
                Combine output records=0
                Reduce input groups=12
                Reduce shuffle bytes=59082
                Reduce input records=6564
                Reduce output records=12
                Spilled Records=6564
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=0
                Total committed heap usage (bytes)=526385152
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Output Format Counters
                Bytes Written=81
2025-05-20 14:53:16,290 INFO mapred.LocalJobRunner: Finishing task: attempt_local2143084439_0001_r_000000_0
2025-05-20 14:53:16,291 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-05-20 14:53:16,895 INFO mapreduce.Job: Job job_local2143084439_0001 running in uber mode : false
2025-05-20 14:53:16,897 INFO mapreduce.Job:  map 100% reduce 100%
2025-05-20 14:53:16,899 INFO mapreduce.Job: Job job_local2143084439_0001 completed successfully
2025-05-20 14:53:16,908 INFO mapreduce.Job: Counters: 36
        File System Counters
                FILE: Number of bytes read=127342
                FILE: Number of bytes written=1467304
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1776380
                HDFS: Number of bytes written=81
                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=6565
                Map output records=6564
                Map output bytes=45948
                Map output materialized bytes=59082
                Input split bytes=95
```


hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
FILE: Number of bytes written=1467304
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1776380
HDFS: Number of bytes written=81
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=6565
Map output records=6564
Map output bytes=45948
Map output materialized bytes=59082
Input split bytes=95
Combine input records=0
Combine output records=0
Reduce input groups=12
Reduce shuffle bytes=59082
Reduce input records=6564
Reduce output records=12
Spilled Records=13128
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=888190
File Output Format Counters
Bytes Written=81
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /omn/part-r-00000
cat: '/omn/part-r-00000': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /ren/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-800-G9-Desktop-PC:~$


Lab-6
Hadoop
1. $ start-all.sh
Output: Starting namenodes on [localhost]
localhost: starting namenode, logging to ...
Starting datanodes
localhost: starting datanode, logging to ...
2. $ hdfs dfs -mkdir /bda_hadoop
3. $ hadoop fs -ls /
Output: Found 1 items
4. $ hdfs dfs -put /home/hduser/Desktop/bda_local.txt bda_hadoop/file.txt
5. $ hdfs dfs -cat /bda_hadoop/file.txt
Output: Hello Prathanjan
-> Copy from local
6. $ hdfs dfs -copyFromLocal /home/hduser/Desktop/bda_local.txt bda_hadoop/file_cp_local.txt
7. $ hdfs dfs -cat /bda_hadoop/file_cp_local.txt
Output: Hello Prathanjan



8. $ hdfs dfs -get /bda_hadoop/file.txt /home/hduser/Downloaded_file.txt

9. $ hdfs dfs -getmerge /bda_hadoop/file.txt /bda_hadoop/file_cp_local.txt /home/hduser/Downloads/

10. $ hadoop fs -getfacl /bda_hadoop/

Output: #file: /bda_hadoop
owner: hduser
group: supergroup
user::rwx
group::r-x
other::r-x

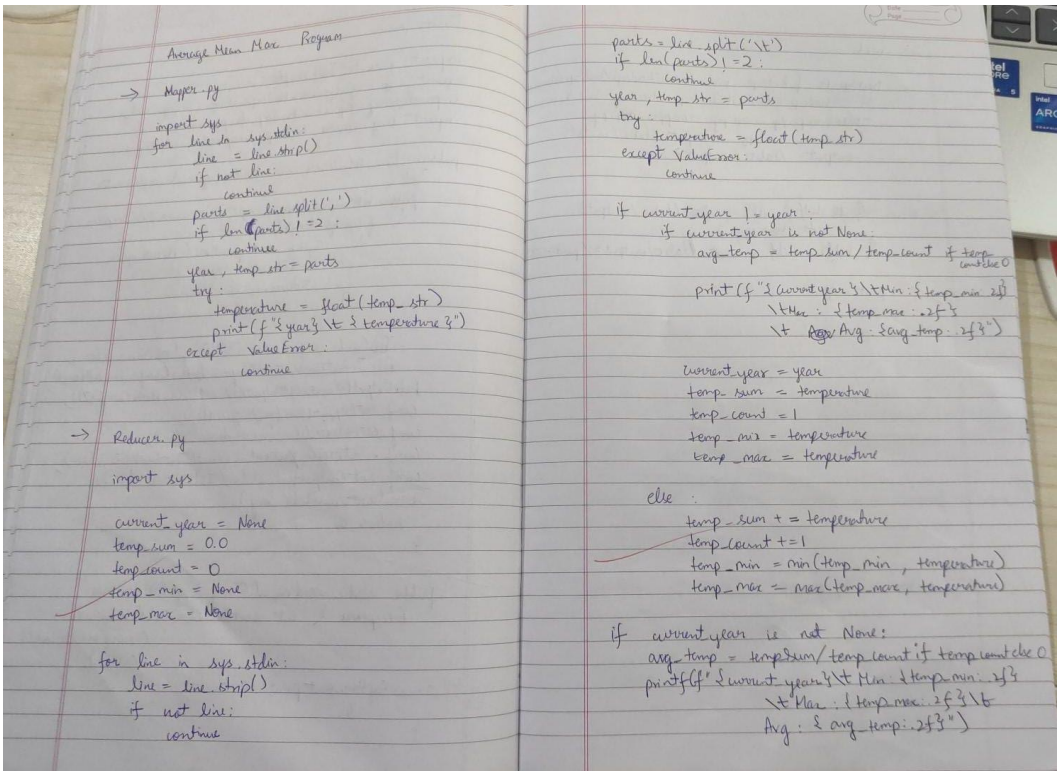
11. $ hdfs dfs -copyToLocal /bda_hadoop/file.txt /home/hduser/Desktop

12. $ hadoop fs -mv /bda_hadoop /abc

13. $ hadoop fs -ls /abc

Output: Found 1 items
drwxr-xr-x - hduser supergroup 0 4444-MM-HH HH:MM /abc/bda_hadoop

14. $ hadoop fs -cp /hello/hadoop_lab



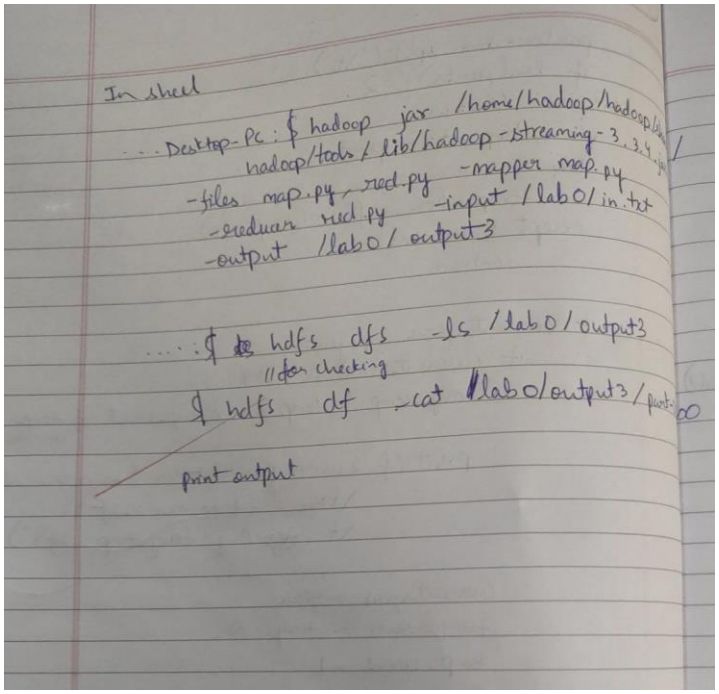
Average Mean Max Program

-> Mapper.py

```
import sys
for line in sys.stdin:
    line = line.strip()
    if not line:
        continue
    parts = line.split(',')
    if len(parts) != 2:
        continue
    year, temp_str = parts
    try:
        temperature = float(temp_str)
        print(f"{year}\t{temperature}")
    except ValueError:
        continue
```

-> Reducer.py

```
import sys

current_year = None
temp_sum = 0.0
temp_count = 0
temp_min = None
temp_max = None

for line in sys.stdin:
    line = line.strip()
    if not line:
        continue
    parts = line.split('\t')
    if len(parts) != 2:
        continue
    year, temp_str = parts
    try:
        temperature = float(temp_str)
    except ValueError:
        continue

    if current_year != year:
        if current_year is not None:
            avg_temp = temp_sum / temp_count if temp_count else 0
            print(f"{current_year}\tMin: {temp_min:.2f}
                \tMax: {temp_max:.2f}
                \tAvg: {avg_temp:.2f}")

        current_year = year
        temp_sum = temperature
        temp_count = 1
        temp_min = temperature
        temp_max = temperature

    else:
        temp_sum += temperature
        temp_count += 1
        temp_min = min(temp_min, temperature)
        temp_max = max(temp_max, temperature)

if current_year is not None:
    avg_temp = temp_sum/temp_count if temp_count else 0
    print(f"{current_year}\tMin: {temp_min:.2f}\t
        Max: {temp_max:.2f}\t
        Avg: {avg_temp:.2f}")
```



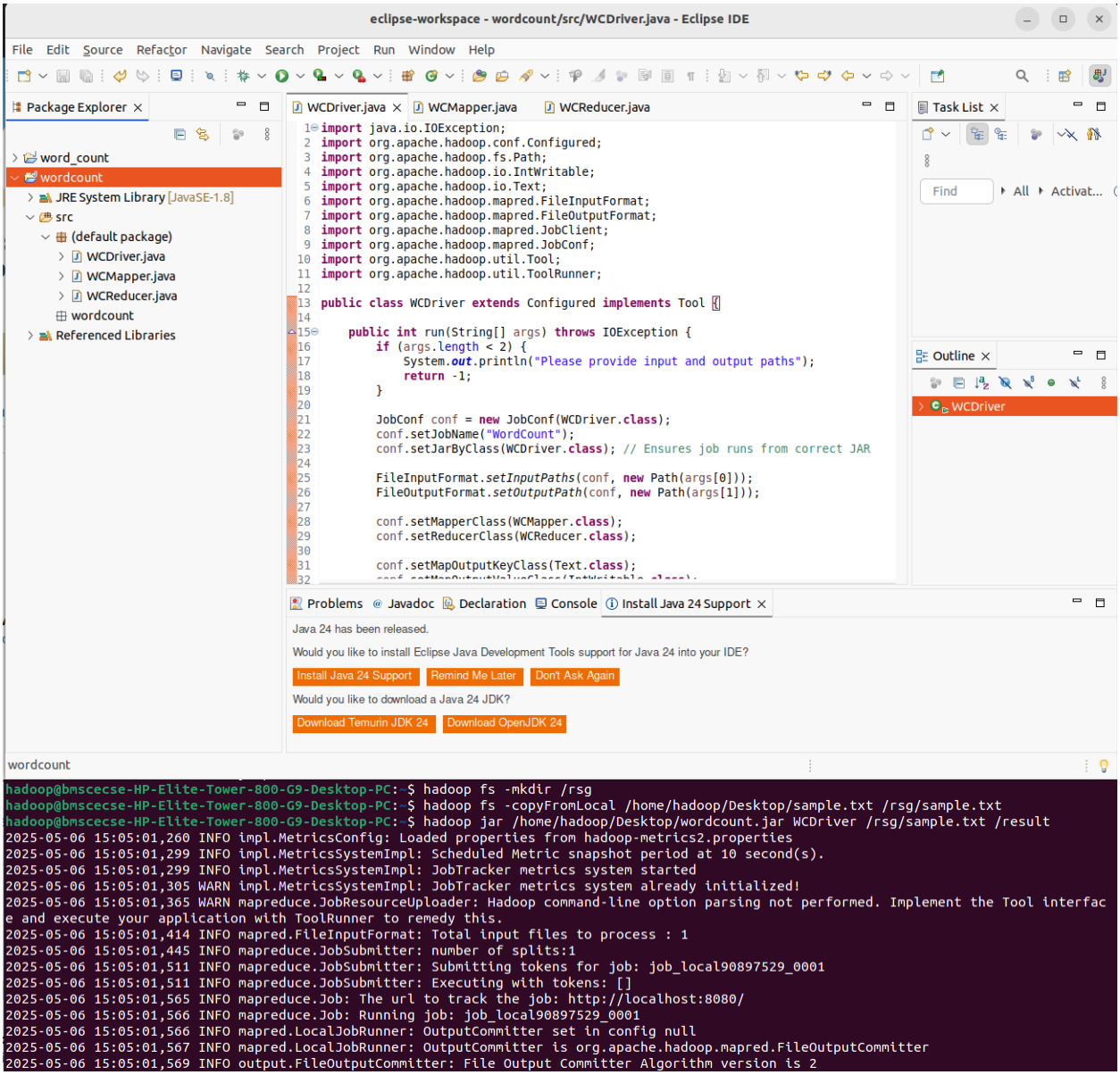
In shell

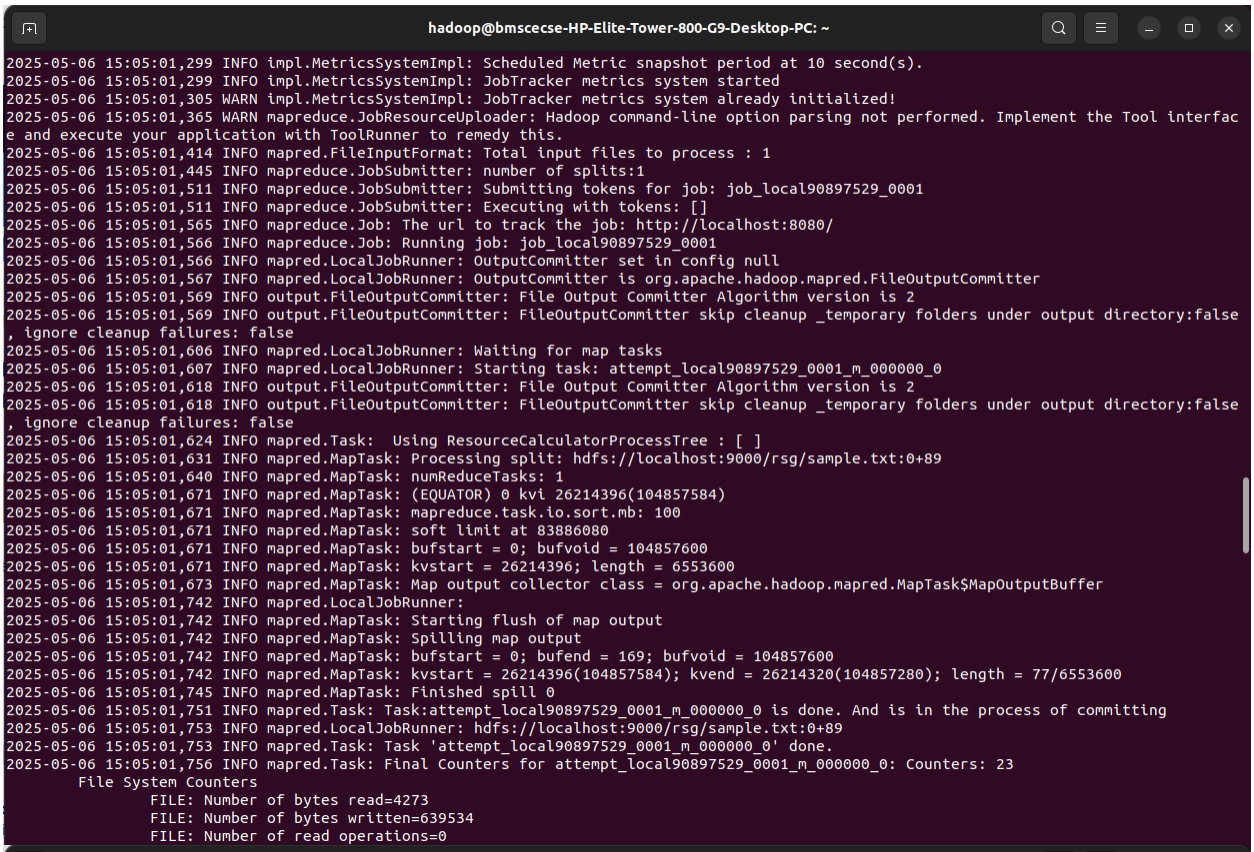
```
...Desktop-Pc: $ hadoop jar /home/hadoop/hadoop/lib/
    hadoop/tools/lib/hadoop-streaming-3.3.4.jar
    -files map.py,red.py -mapper map.py
    -reducer red.py -input /lab0/in.txt
    -output /lab0/output3

....: $ hdfs dfs -ls /lab0/output3
    // for checking
$ hdfs df -cat /lab0/output3/part-00

print output
```

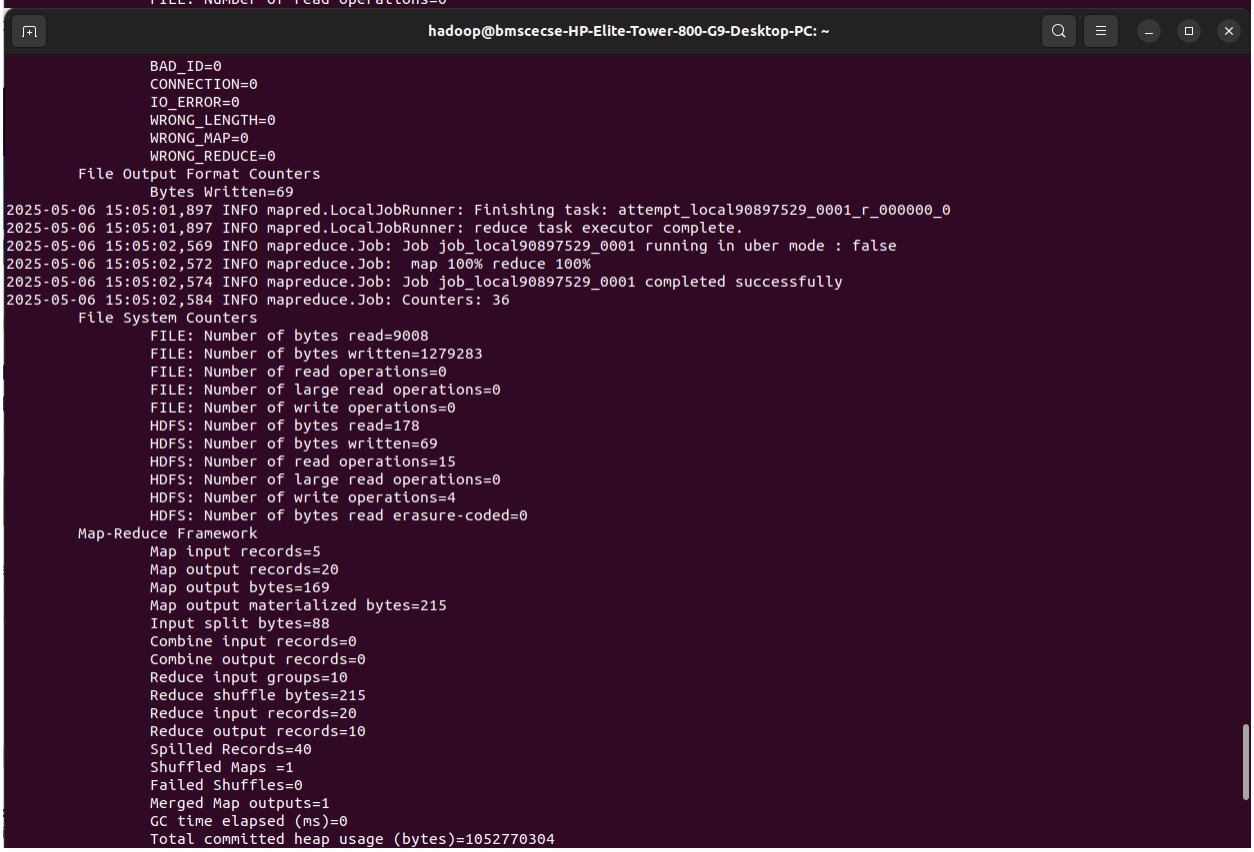
Lab 6: Hadoop

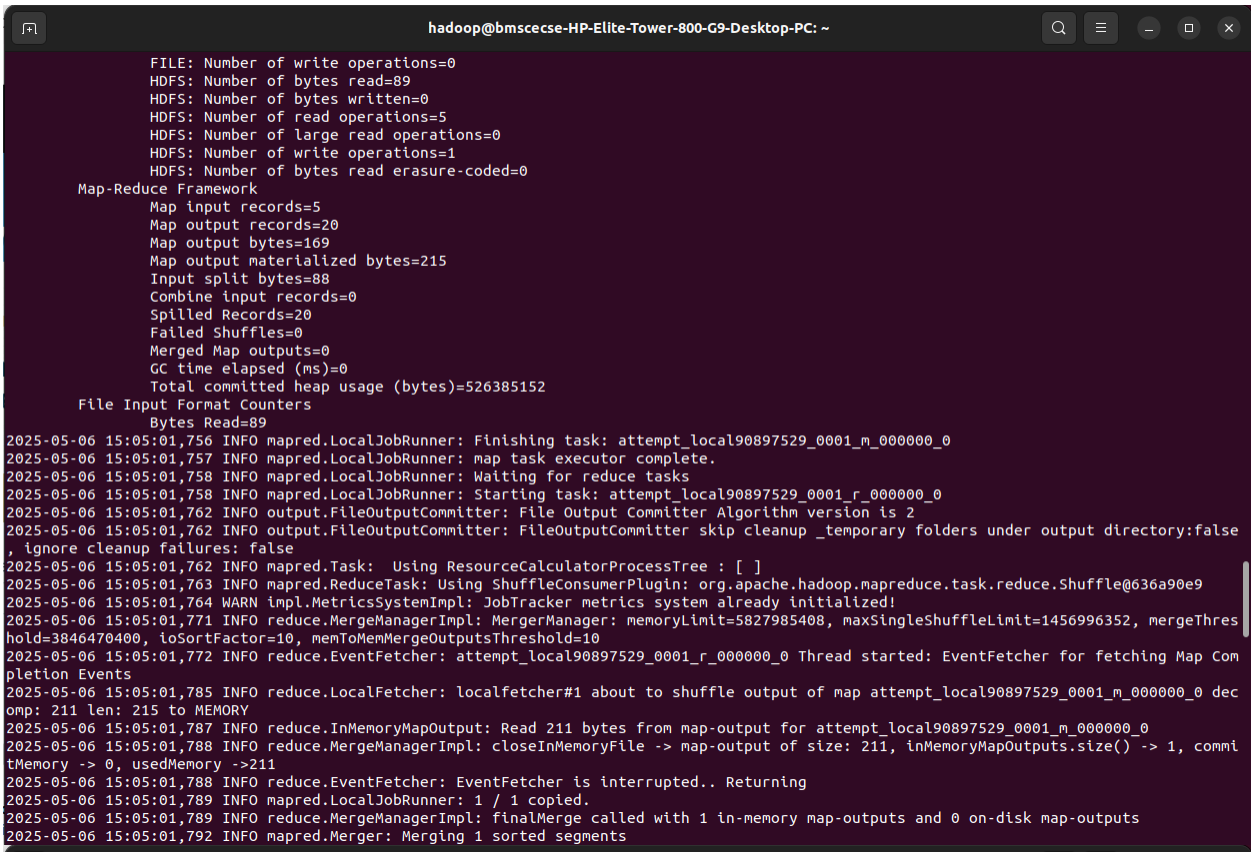
Question: Implement Wordcount program on Hadoop framework

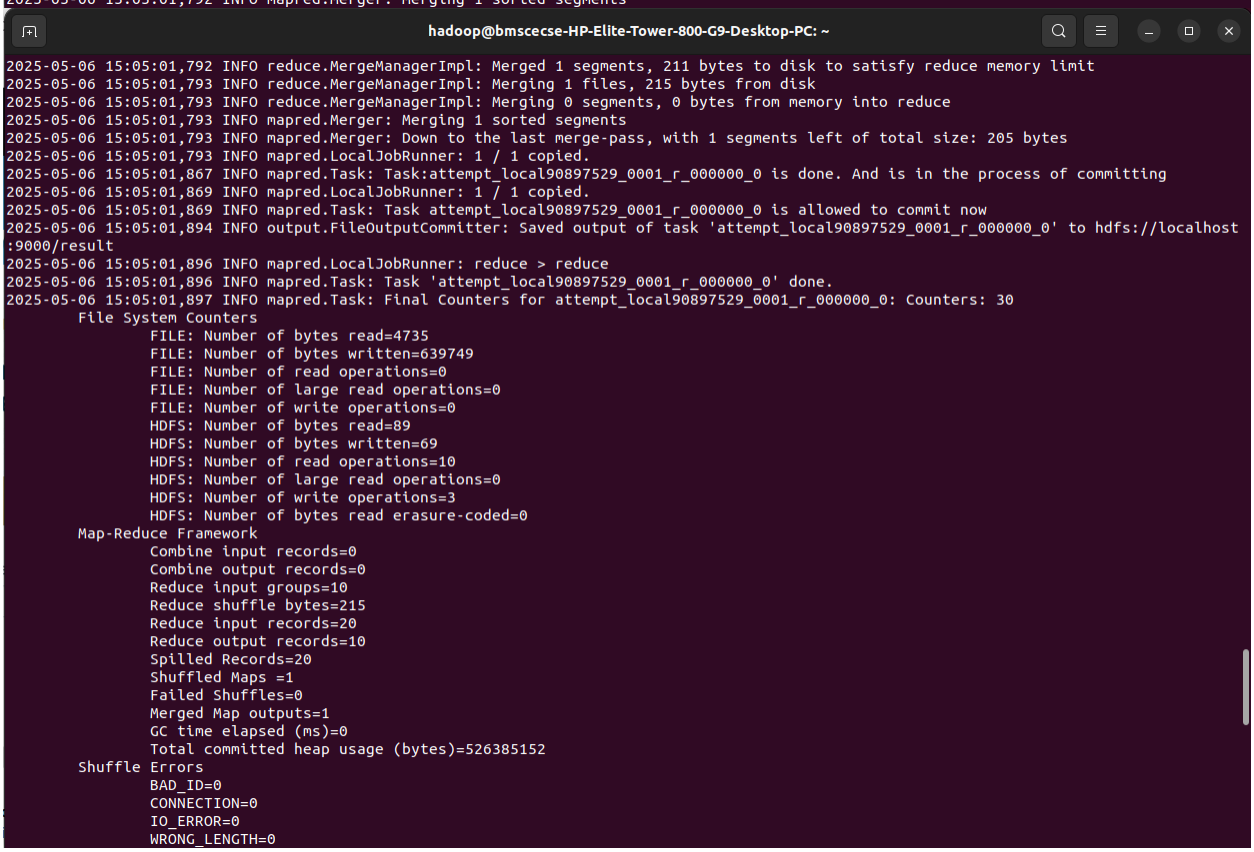
Code with Output:



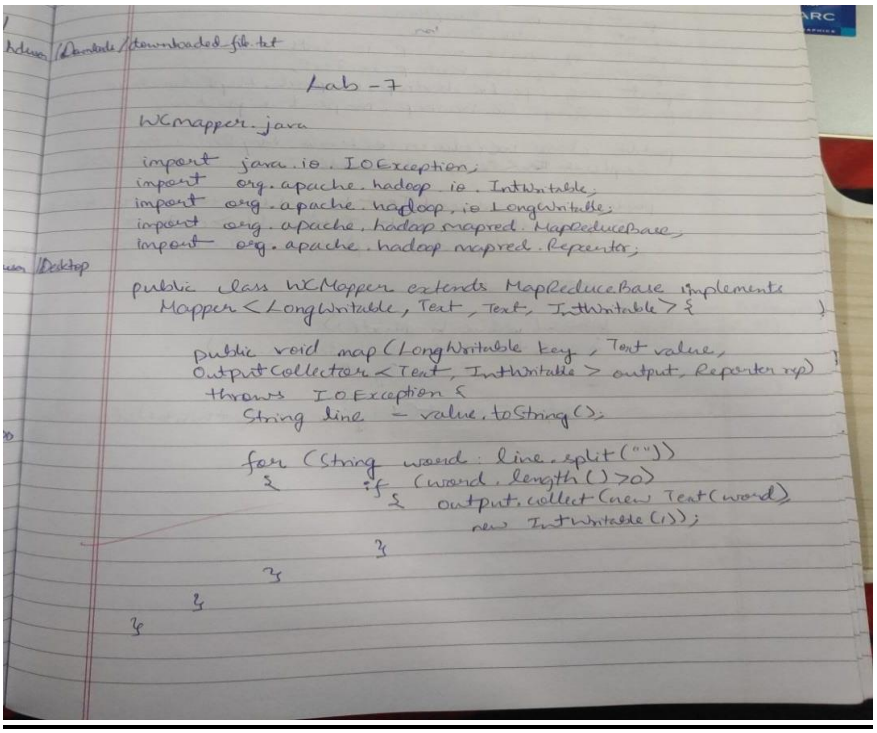

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
2025-05-06 15:05:01,299 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-05-06 15:05:01,299 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-05-06 15:05:01,305 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-05-06 15:05:01,365 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interfac
e and execute your application with ToolRunner to remedy this.
2025-05-06 15:05:01,414 INFO mapred.FileInputFormat: Total input files to process : 1
2025-05-06 15:05:01,445 INFO mapreduce.JobSubmitter: number of splits:1
2025-05-06 15:05:01,511 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local90897529_0001
2025-05-06 15:05:01,511 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-06 15:05:01,565 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-05-06 15:05:01,566 INFO mapreduce.Job: Running job: job_local90897529_0001
2025-05-06 15:05:01,566 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-05-06 15:05:01,567 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2025-05-06 15:05:01,569 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-06 15:05:01,569 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false
, ignore cleanup failures: false
2025-05-06 15:05:01,606 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-05-06 15:05:01,607 INFO mapred.LocalJobRunner: Starting task: attempt_local90897529_0001_m_000000_0
2025-05-06 15:05:01,618 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-06 15:05:01,618 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false
, ignore cleanup failures: false
2025-05-06 15:05:01,624 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2025-05-06 15:05:01,631 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/rsg/sample.txt:0+89
2025-05-06 15:05:01,640 INFO mapred.MapTask: numReduceTasks: 1
2025-05-06 15:05:01,671 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-05-06 15:05:01,671 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-05-06 15:05:01,671 INFO mapred.MapTask: soft limit at 83886080
2025-05-06 15:05:01,671 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-05-06 15:05:01,671 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-05-06 15:05:01,673 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-06 15:05:01,742 INFO mapred.LocalJobRunner:
2025-05-06 15:05:01,742 INFO mapred.MapTask: Starting flush of map output
2025-05-06 15:05:01,742 INFO mapred.MapTask: Spilling map output
2025-05-06 15:05:01,742 INFO mapred.MapTask: bufstart = 0; bufend = 169; bufvoid = 104857600
2025-05-06 15:05:01,742 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214320(104857280); length = 77/6553600
2025-05-06 15:05:01,745 INFO mapred.MapTask: Finished spill 0
2025-05-06 15:05:01,751 INFO mapred.Task: Task:attempt_local90897529_0001_m_000000_0 is done. And is in the process of committing
2025-05-06 15:05:01,753 INFO mapred.LocalJobRunner: hdfs://localhost:9000/rsg/sample.txt:0+89
2025-05-06 15:05:01,753 INFO mapred.Task: Task 'attempt_local90897529_0001_m_000000_0' done.
2025-05-06 15:05:01,756 INFO mapred.Task: Final Counters for attempt_local90897529_0001_m_000000_0: Counters: 23
File System Counters
FILE: Number of bytes read=4273
FILE: Number of bytes written=639534
FILE: Number of read operations=0


hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Output Format Counters
Bytes Written=69
2025-05-06 15:05:01,897 INFO mapred.LocalJobRunner: Finishing task: attempt_local90897529_0001_r_000000_0
2025-05-06 15:05:01,897 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-05-06 15:05:02,569 INFO mapreduce.Job: Job job_local90897529_0001 running in uber mode : false
2025-05-06 15:05:02,572 INFO mapreduce.Job: map 100% reduce 100%
2025-05-06 15:05:02,574 INFO mapreduce.Job: Job job_local90897529_0001 completed successfully
2025-05-06 15:05:02,584 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=9008
FILE: Number of bytes written=1279283
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=178
HDFS: Number of bytes written=69
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=5
Map output records=20
Map output bytes=169
Map output materialized bytes=215
Input split bytes=88
Combine input records=0
Combine output records=0
Reduce input groups=10
Reduce shuffle bytes=215
Reduce input records=20
Reduce output records=10
Spilled Records=40
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=1052770304


hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
FILE: Number of write operations=0
HDFS: Number of bytes read=89
HDFS: Number of bytes written=0
HDFS: Number of read operations=5
HDFS: Number of large read operations=0
HDFS: Number of write operations=1
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=5
Map output records=20
Map output bytes=169
Map output materialized bytes=215
Input split bytes=88
Combine input records=0
Spilled Records=20
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=0
Total committed heap usage (bytes)=526385152
File Input Format Counters
Bytes Read=89
2025-05-06 15:05:01,756 INFO mapred.LocalJobRunner: Finishing task: attempt_local90897529_0001_m_000000_0
2025-05-06 15:05:01,757 INFO mapred.LocalJobRunner: map task executor complete.
2025-05-06 15:05:01,758 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2025-05-06 15:05:01,758 INFO mapred.LocalJobRunner: Starting task: attempt_local90897529_0001_r_000000_0
2025-05-06 15:05:01,762 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-06 15:05:01,762 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false
, ignore cleanup failures: false
2025-05-06 15:05:01,762 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2025-05-06 15:05:01,763 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@636a90e9
2025-05-06 15:05:01,764 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-05-06 15:05:01,771 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=5827985408, maxSingleShuffleLimit=1456996352, mergeThres
hold=3846470400, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-05-06 15:05:01,772 INFO reduce.EventFetcher: attempt_local90897529_0001_r_000000_0 Thread started: EventFetcher for fetching Map Com
pletion Events
2025-05-06 15:05:01,785 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local90897529_0001_m_000000_0 dec
omp: 211 len: 215 to MEMORY
2025-05-06 15:05:01,787 INFO reduce.InMemoryMapOutput: Read 211 bytes from map-output for attempt_local90897529_0001_m_000000_0
2025-05-06 15:05:01,788 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 211, inMemoryMapOutputs.size() -> 1, commi
tMemory -> 0, usedMemory ->211
2025-05-06 15:05:01,788 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2025-05-06 15:05:01,789 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-06 15:05:01,789 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs
2025-05-06 15:05:01,792 INFO mapred.Merger: Merging 1 sorted segments


hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
2025-05-06 15:05:01,792 INFO reduce.MergeManagerImpl: Merged 1 segments, 211 bytes to disk to satisfy reduce memory limit
2025-05-06 15:05:01,793 INFO reduce.MergeManagerImpl: Merging 1 files, 215 bytes from disk
2025-05-06 15:05:01,793 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2025-05-06 15:05:01,793 INFO mapred.Merger: Merging 1 sorted segments
2025-05-06 15:05:01,793 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 205 bytes
2025-05-06 15:05:01,793 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-06 15:05:01,867 INFO mapred.Task: Task:attempt_local90897529_0001_r_000000_0 is done. And is in the process of committing
2025-05-06 15:05:01,869 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-06 15:05:01,869 INFO mapred.Task: Task attempt_local90897529_0001_r_000000_0 is allowed to commit now
2025-05-06 15:05:01,894 INFO output.FileOutputCommitter: Saved output of task 'attempt_local90897529_0001_r_000000_0' to hdfs://localhost
:9000/result
2025-05-06 15:05:01,896 INFO mapred.LocalJobRunner: reduce > reduce
2025-05-06 15:05:01,896 INFO mapred.Task: Task 'attempt_local90897529_0001_r_000000_0' done.
2025-05-06 15:05:01,897 INFO mapred.Task: Final Counters for attempt_local90897529_0001_r_000000_0: Counters: 30
File System Counters
FILE: Number of bytes read=4735
FILE: Number of bytes written=639749
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=89
HDFS: Number of bytes written=69
HDFS: Number of read operations=10
HDFS: Number of large read operations=0
HDFS: Number of write operations=3
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Combine input records=0
Combine output records=0
Reduce input groups=10
Reduce shuffle bytes=215
Reduce input records=20
Reduce output records=10
Spilled Records=20
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=526385152
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0

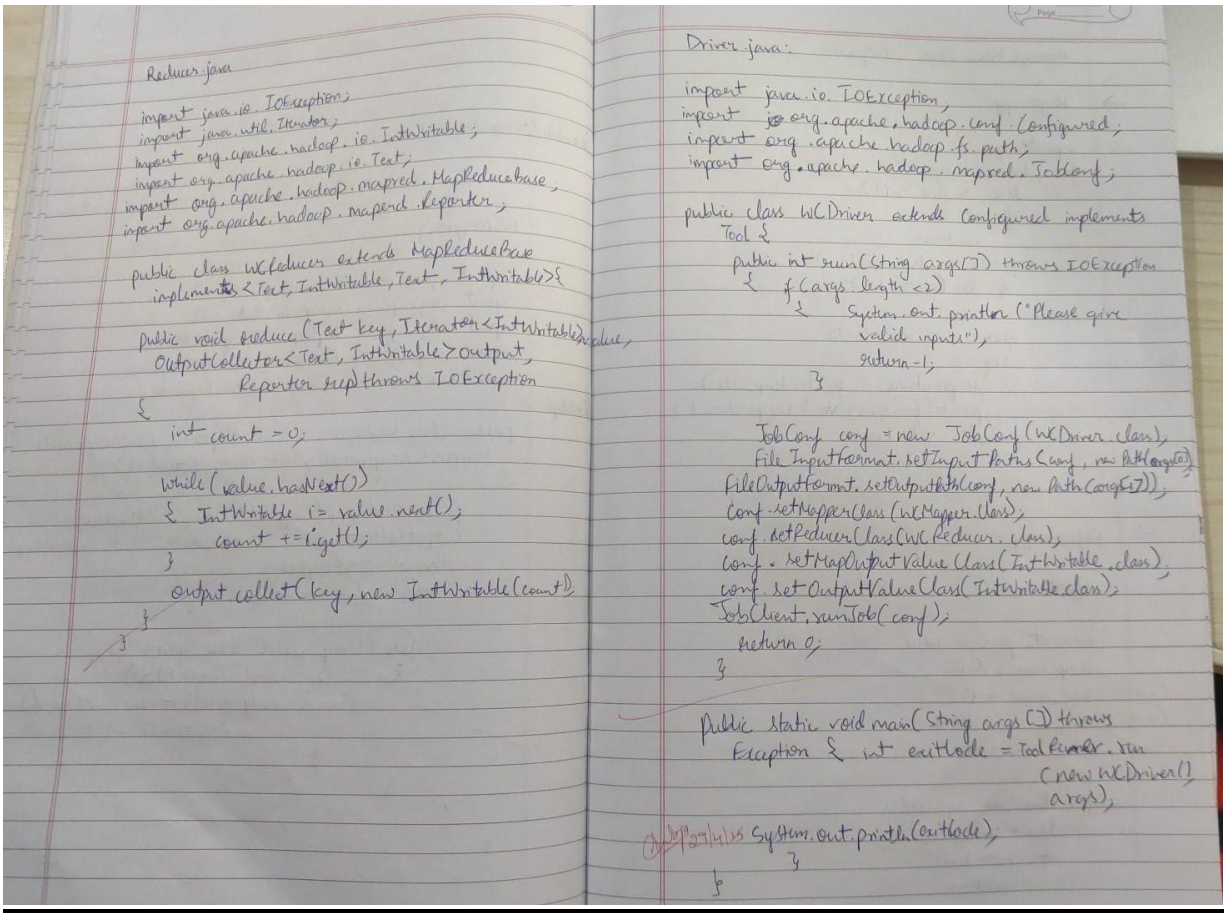
```
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
                HDFS: Number of write operations=4
                HDFS: Number of bytes read erasure-coded=0
        Map-Reduce Framework
                Map input records=5
                Map output records=20
                Map output bytes=169
                Map output materialized bytes=215
                Input split bytes=88
                Combine input records=0
                Combine output records=0
                Reduce input groups=10
                Reduce shuffle bytes=215
                Reduce input records=20
                Reduce output records=10
                Spilled Records=40
                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=89
        File Output Format Counters
                Bytes Written=69
Exit Code: 0
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /result/part-00000
are     1
brother 1
family  1
hi      1
how     5
is      4
job     1
sister  1
you     1
your    4
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$
```



Lab - 7

WCmapper.java

```
import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.Reporter;

public class WCMapper extends MapReduceBase implements
  Mapper<LongWritable, Text, Text, IntWritable> {

  public void map(LongWritable key, Text value,
  OutputCollector<Text, IntWritable> output, Reporter rp)
  throws IOException {
    String line = value.toString();

    for (String word: line.split(" "))
    {
      if (word.length()>0)
      {
        output.collect(new Text(word),
                new IntWritable(1));
      }
    }
  }
}
```



Reducer.java

```
import java.io.IOException;
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.Reporter,

public class WCReducer extends MapReduceBase
  implements <Text, IntWritable, Text, IntWritable>{

public void reduce (Text key, Iterator<IntWritable> value,
  OutputCollector<Text, IntWritable> output,
        Reporter rep) throws IOException
{
  int count = 0;

  While (value.hasNext())
  { IntWritable i = value.next();
    count += i.get();
  }
  output.collect(key, new IntWritable(count));
}
}
```

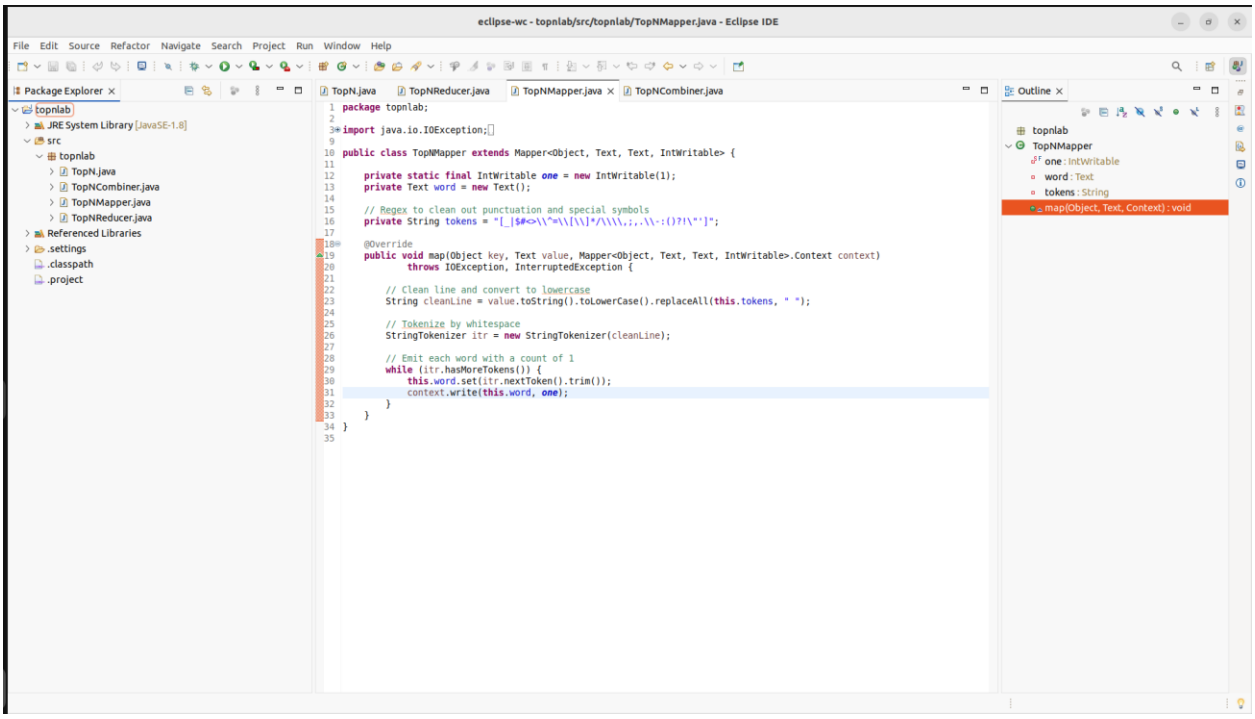
Driver.java:

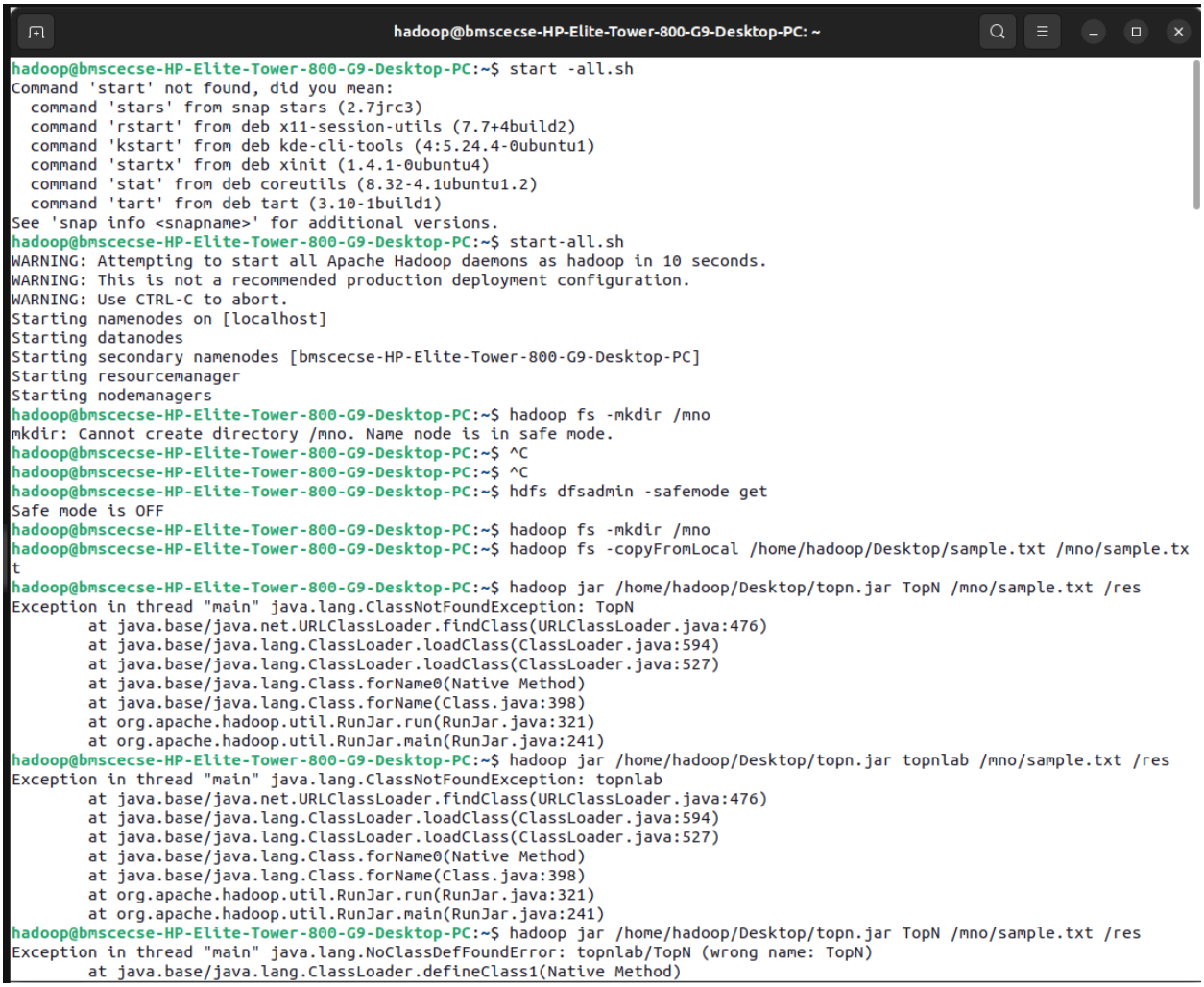
```
import java.io.IOException,
import org.apache.hadoop.conf.Configured,
import org.apache.hadoop.fs.path;
import org.apache.hadoop.mapred.JobConf,

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.setReducer(class(WCReducer.class),
conf.setMapOutputValue Class(IntWritable.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),
    }
}
```

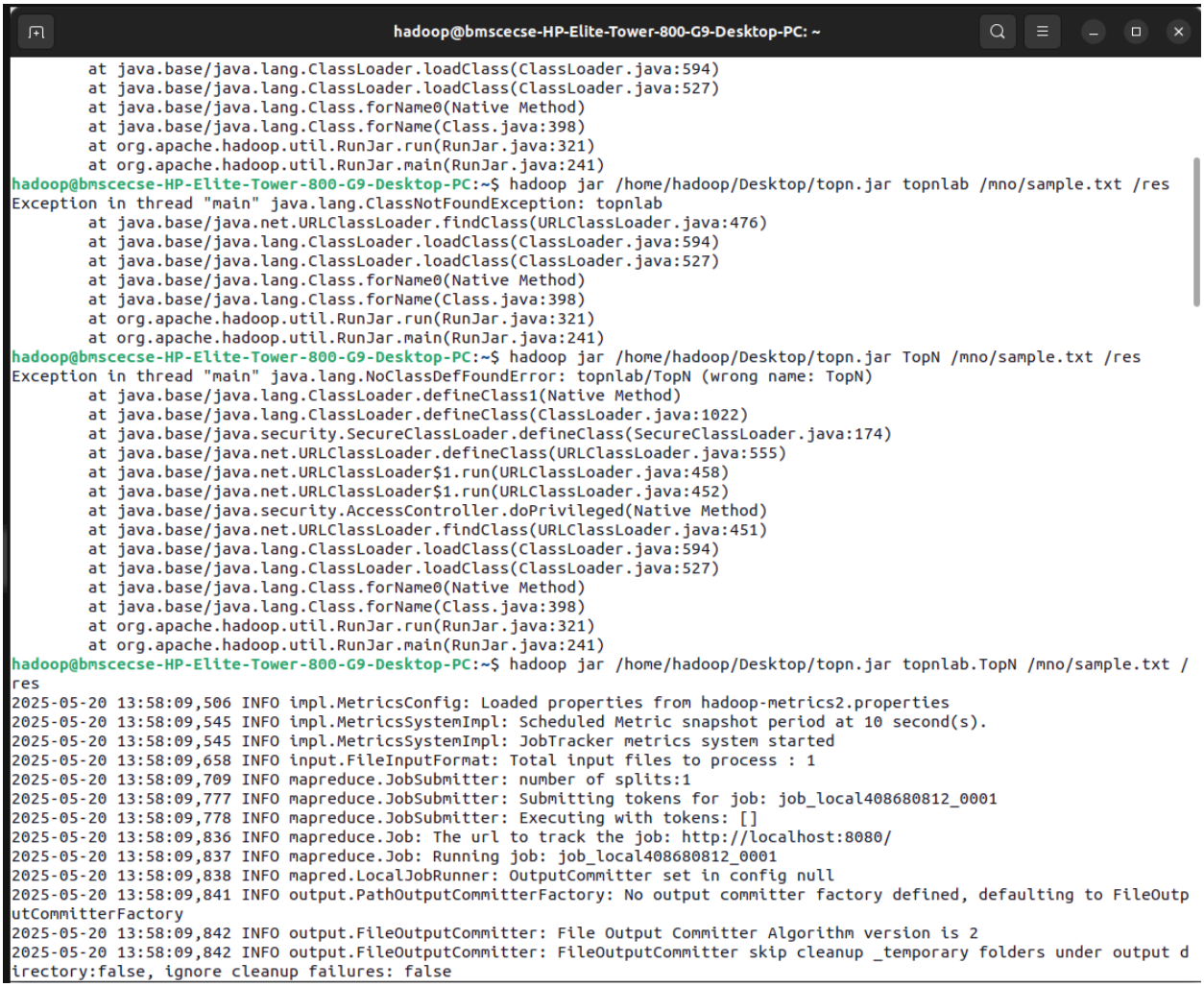
Lab 7: Hadoop

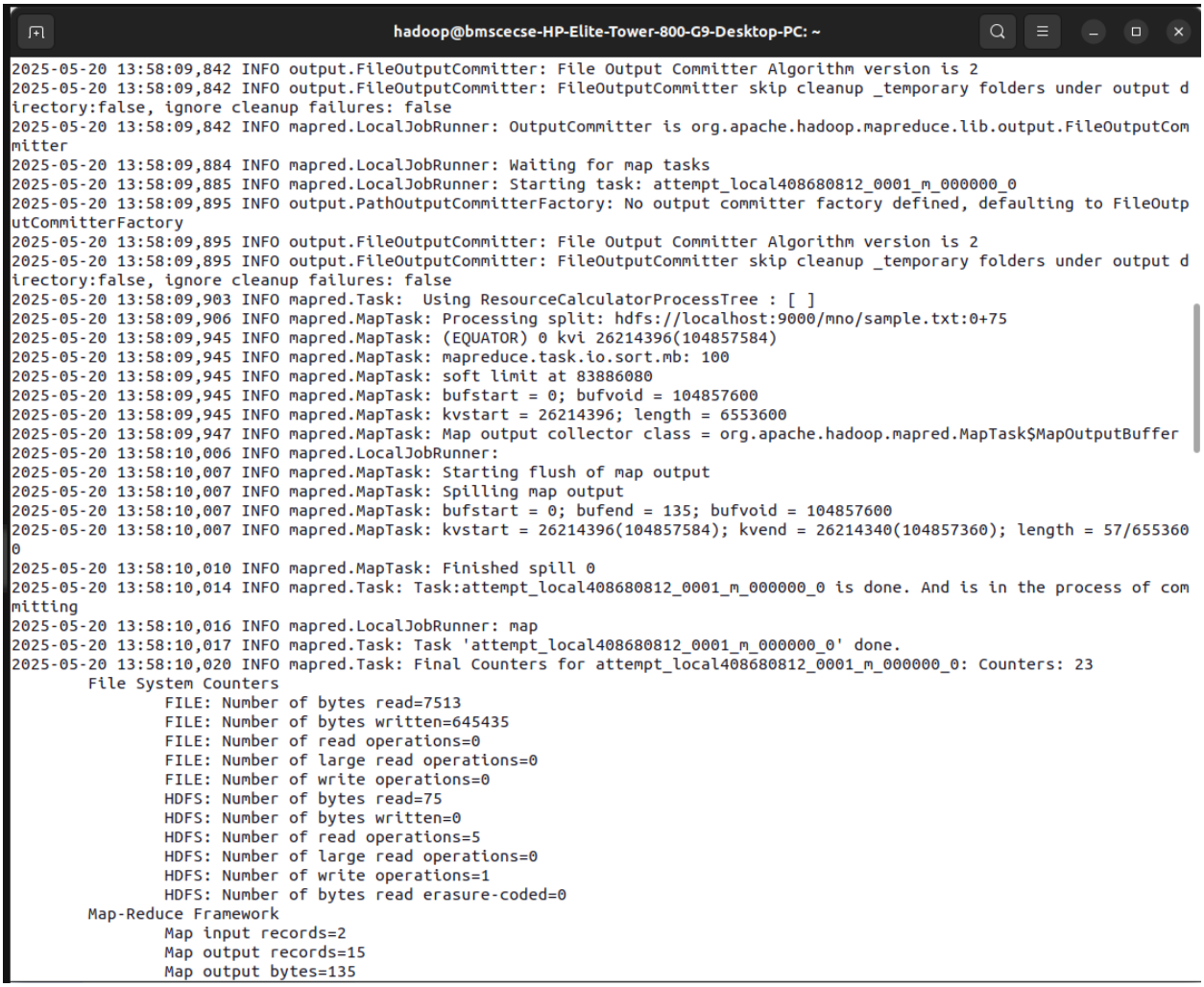
Question: For a given Text file, Create a Map Reduce program to sort the content in an alphabetic order listing only top 10 maximum occurrences of words.

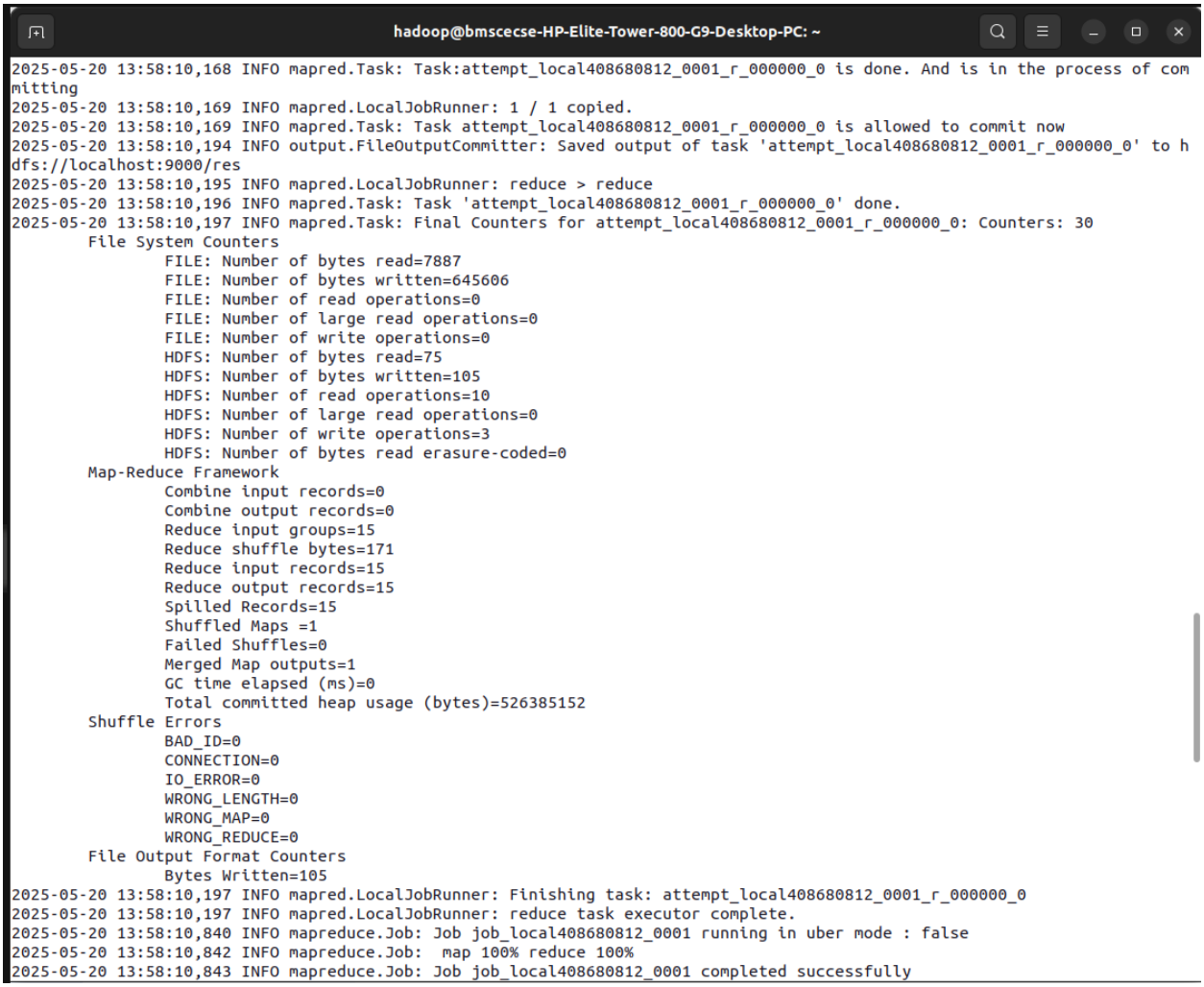
Code with Output:

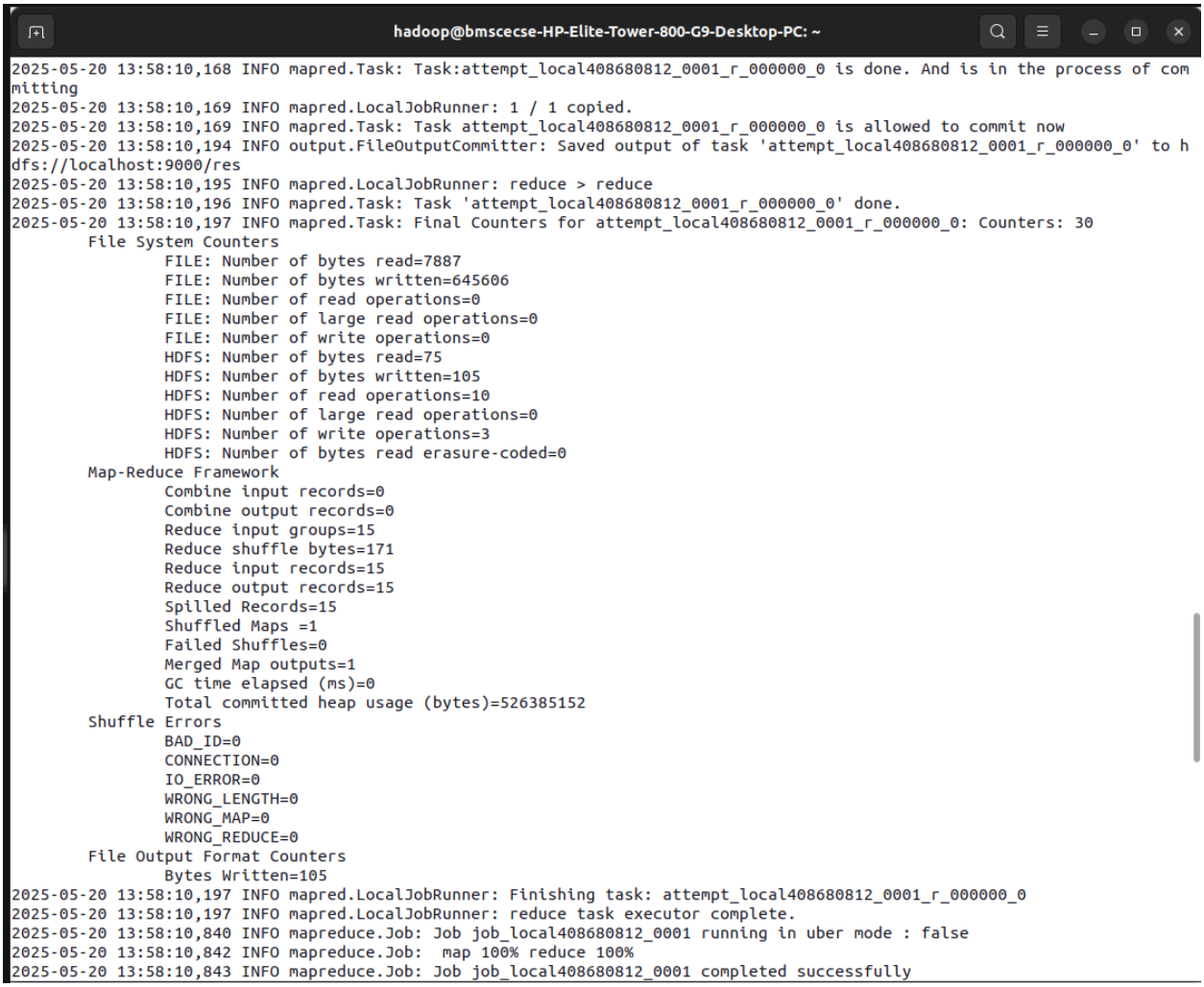


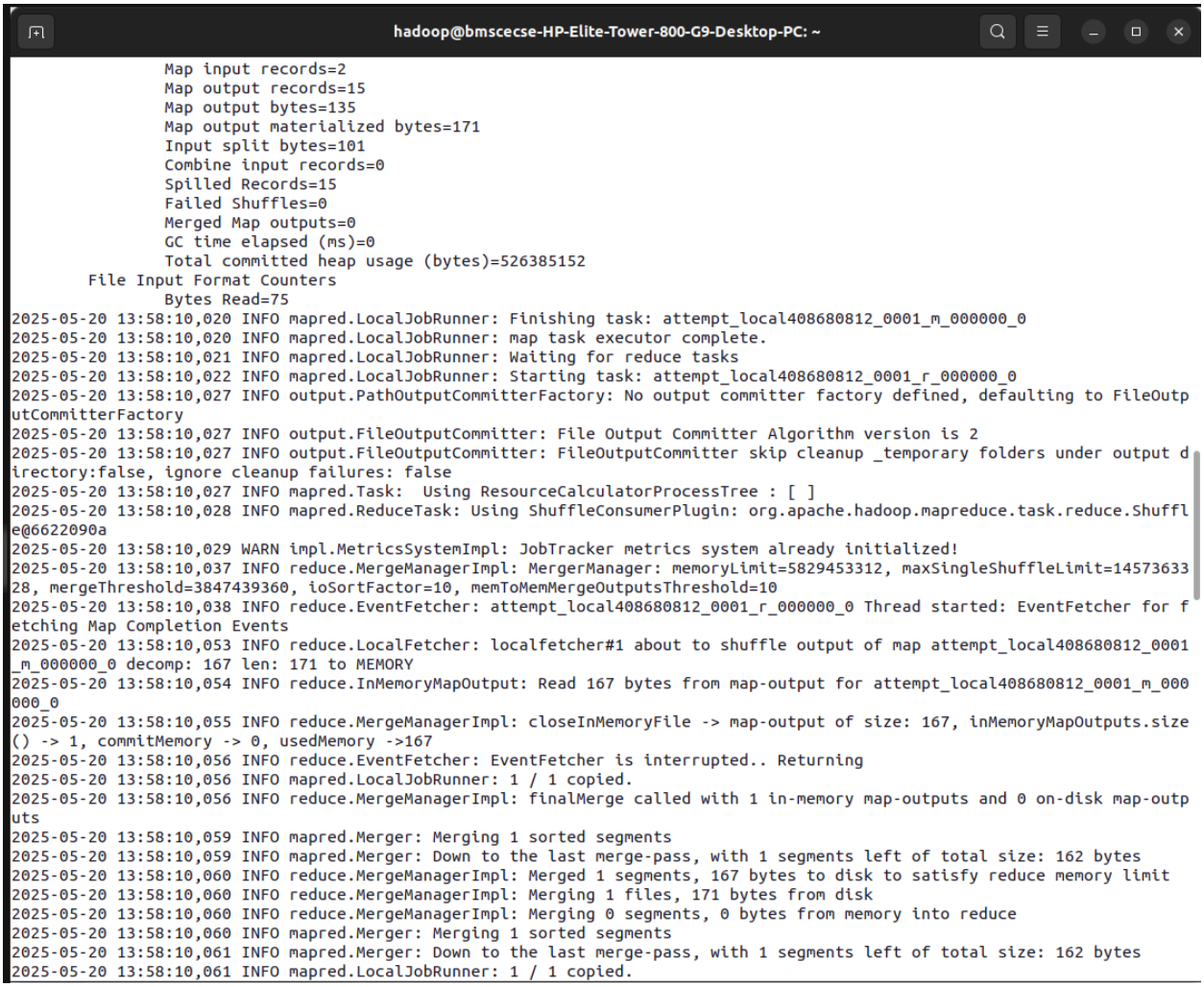

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ start -all.sh
Command 'start' not found, did you mean:
command 'stars' from snap stars (2.7jrc3)
command 'rstart' from deb x11-session-utils (7.7+4build2)
command 'kstart' from deb kde-cli-tools (4:5.24.4-0ubuntu1)
command 'startx' from deb xinit (1.4.1-0ubuntu4)
command 'stat' from deb coreutils (8.32-4.1ubuntu1.2)
command 'tart' from deb tart (3.10-1build1)
See 'snap info <snapname>' for additional versions.
hadoop@bmscecse-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 [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
Starting resourcemanager
Starting nodemanagers
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mkdir /mno
mkdir: Cannot create directory /mno. Name node is in safe mode.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ ^C
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ ^C
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfsadmin -safemode get
Safe mode is OFF
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mkdir /mno
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -copyFromLocal /home/hadoop/Desktop/sample.txt /mno/sample.tx
t
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/topn.jar TopN /mno/sample.txt /res
Exception in thread "main" java.lang.ClassNotFoundException: TopN
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-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/topn.jar topnlab /mno/sample.txt /res
Exception in thread "main" java.lang.ClassNotFoundException: topnlab
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-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/topn.jar TopN /mno/sample.txt /res
Exception in thread "main" java.lang.NoClassDefFoundError: topnlab/TopN (wrong name: TopN)
at java.base/java.lang.ClassLoader.defineClass1(Native Method)

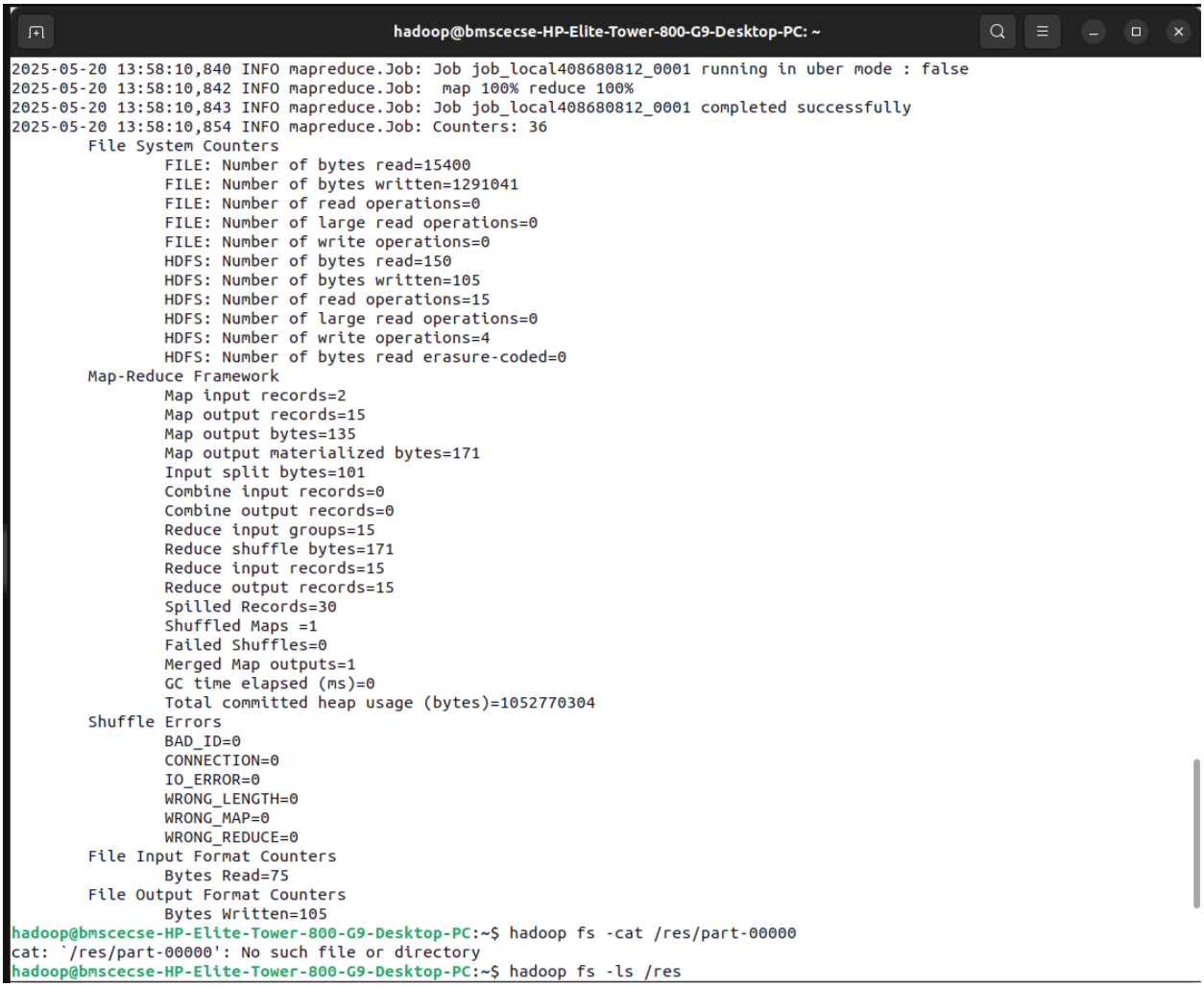

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:594)
at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:527)
at java.base/java.lang.Class.forName0(Native Method)
at java.base/java.lang.Class.forName(Class.java:398)
at org.apache.hadoop.util.RunJar.run(RunJar.java:321)
at org.apache.hadoop.util.RunJar.main(RunJar.java:241)
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/topn.jar topnlab /mno/sample.txt /res
Exception in thread "main" java.lang.ClassNotFoundException: topnlab
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-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/topn.jar TopN /mno/sample.txt /res
Exception in thread "main" java.lang.NoClassDefFoundError: topnlab/TopN (wrong name: TopN)
at java.base/java.lang.ClassLoader.defineClass1(Native Method)
at java.base/java.lang.ClassLoader.defineClass(ClassLoader.java:1022)
at java.base/java.security.SecureClassLoader.defineClass(SecureClassLoader.java:174)
at java.base/java.net.URLClassLoader.defineClass(URLClassLoader.java:555)
at java.base/java.net.URLClassLoader$1.run(URLClassLoader.java:458)
at java.base/java.net.URLClassLoader$1.run(URLClassLoader.java:452)
at java.base/java.security.AccessController.doPrivileged(Native Method)
at java.base/java.net.URLClassLoader.findClass(URLClassLoader.java:451)
at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:594)
at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:527)
at java.base/java.lang.Class.forName0(Native Method)
at java.base/java.lang.Class.forName(Class.java:398)
at org.apache.hadoop.util.RunJar.run(RunJar.java:321)
at org.apache.hadoop.util.RunJar.main(RunJar.java:241)
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/topn.jar topnlab.TopN /mno/sample.txt /
res
2025-05-20 13:58:09,506 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-05-20 13:58:09,545 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-05-20 13:58:09,545 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-05-20 13:58:09,658 INFO input.FileInputFormat: Total input files to process : 1
2025-05-20 13:58:09,709 INFO mapreduce.JobSubmitter: number of splits:1
2025-05-20 13:58:09,777 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local408680812_0001
2025-05-20 13:58:09,778 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-20 13:58:09,836 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-05-20 13:58:09,837 INFO mapreduce.Job: Running job: job_local408680812_0001
2025-05-20 13:58:09,838 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-05-20 13:58:09,841 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutp
utCommitterFactory
2025-05-20 13:58:09,842 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 13:58:09,842 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output d
irectory:false, ignore cleanup failures: false

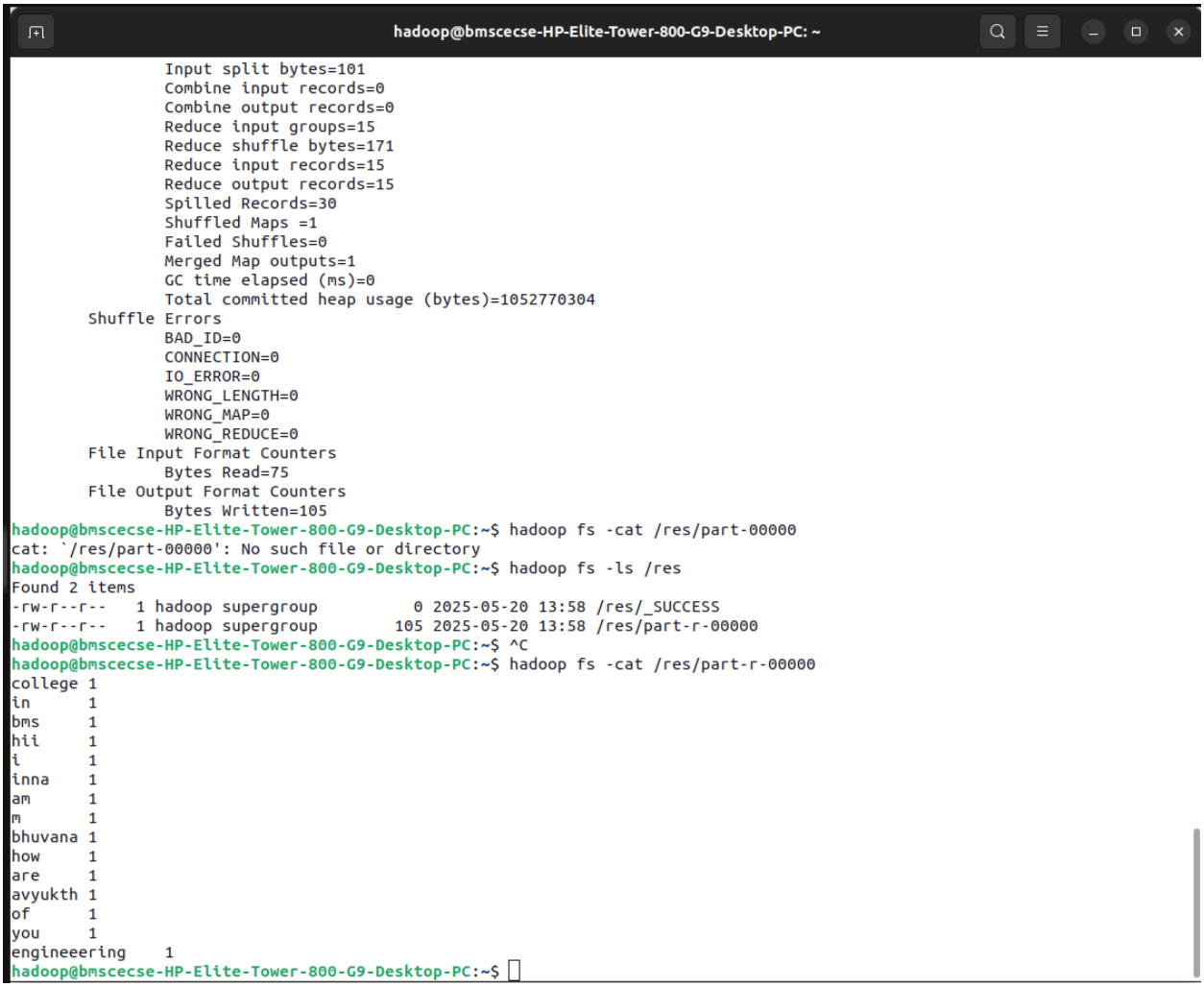

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
2025-05-20 13:58:09,842 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 13:58:09,842 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output d
irectory:false, ignore cleanup failures: false
2025-05-20 13:58:09,842 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCom
mitter
2025-05-20 13:58:09,884 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-05-20 13:58:09,885 INFO mapred.LocalJobRunner: Starting task: attempt_local408680812_0001_m_000000_0
2025-05-20 13:58:09,895 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutp
utCommitterFactory
2025-05-20 13:58:09,895 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 13:58:09,895 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output d
irectory:false, ignore cleanup failures: false
2025-05-20 13:58:09,903 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2025-05-20 13:58:09,906 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/mno/sample.txt:0+75
2025-05-20 13:58:09,945 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-05-20 13:58:09,945 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-05-20 13:58:09,945 INFO mapred.MapTask: soft limit at 83886080
2025-05-20 13:58:09,945 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-05-20 13:58:09,945 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-05-20 13:58:09,947 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-20 13:58:10,006 INFO mapred.LocalJobRunner:
2025-05-20 13:58:10,007 INFO mapred.MapTask: Starting flush of map output
2025-05-20 13:58:10,007 INFO mapred.MapTask: Spilling map output
2025-05-20 13:58:10,007 INFO mapred.MapTask: bufstart = 0; bufend = 135; bufvoid = 104857600
2025-05-20 13:58:10,007 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214340(104857360); length = 57/655360
0
2025-05-20 13:58:10,010 INFO mapred.MapTask: Finished spill 0
2025-05-20 13:58:10,014 INFO mapred.Task: Task:attempt_local408680812_0001_m_000000_0 is done. And is in the process of com
mitting
2025-05-20 13:58:10,016 INFO mapred.LocalJobRunner: map
2025-05-20 13:58:10,017 INFO mapred.Task: Task 'attempt_local408680812_0001_m_000000_0' done.
2025-05-20 13:58:10,020 INFO mapred.Task: Final Counters for attempt_local408680812_0001_m_000000_0: Counters: 23
File System Counters
FILE: Number of bytes read=7513
FILE: Number of bytes written=645435
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=75
HDFS: Number of bytes written=0
HDFS: Number of read operations=5
HDFS: Number of large read operations=0
HDFS: Number of write operations=1
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=2
Map output records=15
Map output bytes=135

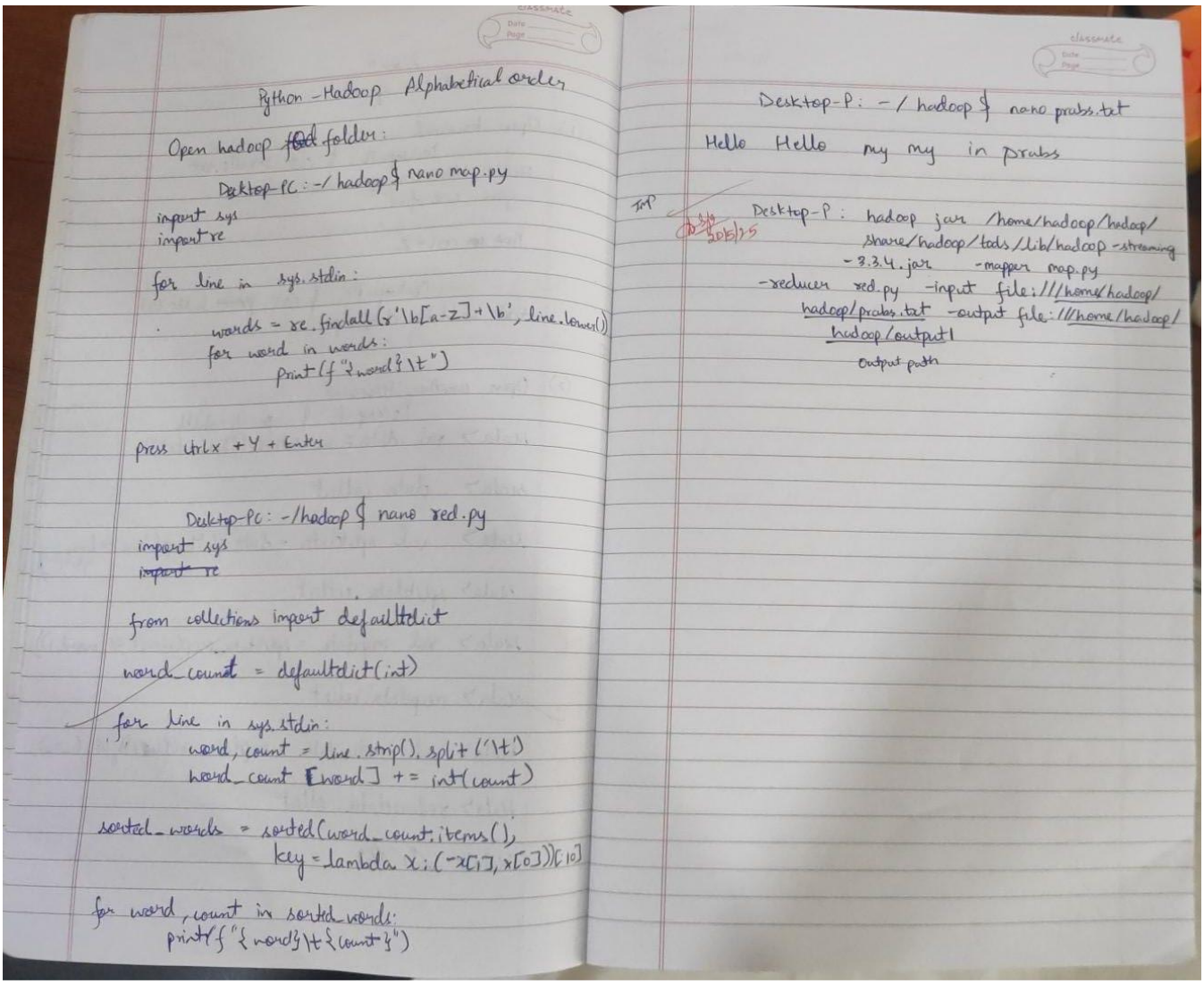

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
2025-05-20 13:58:10,168 INFO mapred.Task: Task:attempt_local408680812_0001_r_000000_0 is done. And is in the process of com
mitting
2025-05-20 13:58:10,169 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 13:58:10,169 INFO mapred.Task: Task attempt_local408680812_0001_r_000000_0 is allowed to commit now
2025-05-20 13:58:10,194 INFO output.FileOutputCommitter: Saved output of task 'attempt_local408680812_0001_r_000000_0' to h
dfs://localhost:9000/res
2025-05-20 13:58:10,195 INFO mapred.LocalJobRunner: reduce > reduce
2025-05-20 13:58:10,196 INFO mapred.Task: Task 'attempt_local408680812_0001_r_000000_0' done.
2025-05-20 13:58:10,197 INFO mapred.Task: Final Counters for attempt_local408680812_0001_r_000000_0: Counters: 30
File System Counters
FILE: Number of bytes read=7887
FILE: Number of bytes written=645606
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=75
HDFS: Number of bytes written=105
HDFS: Number of read operations=10
HDFS: Number of large read operations=0
HDFS: Number of write operations=3
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Combine input records=0
Combine output records=0
Reduce input groups=15
Reduce shuffle bytes=171
Reduce input records=15
Reduce output records=15
Spilled Records=15
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=526385152
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Output Format Counters
Bytes Written=105
2025-05-20 13:58:10,197 INFO mapred.LocalJobRunner: Finishing task: attempt_local408680812_0001_r_000000_0
2025-05-20 13:58:10,197 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-05-20 13:58:10,840 INFO mapreduce.Job: Job job_local408680812_0001 running in uber mode : false
2025-05-20 13:58:10,842 INFO mapreduce.Job: map 100% reduce 100%
2025-05-20 13:58:10,843 INFO mapreduce.Job: Job job_local408680812_0001 completed successfully


hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
2025-05-20 13:58:10,168 INFO mapred.Task: Task:attempt_local408680812_0001_r_000000_0 is done. And is in the process of com
mitting
2025-05-20 13:58:10,169 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 13:58:10,169 INFO mapred.Task: Task attempt_local408680812_0001_r_000000_0 is allowed to commit now
2025-05-20 13:58:10,194 INFO output.FileOutputCommitter: Saved output of task 'attempt_local408680812_0001_r_000000_0' to h
dfs://localhost:9000/res
2025-05-20 13:58:10,195 INFO mapred.LocalJobRunner: reduce > reduce
2025-05-20 13:58:10,196 INFO mapred.Task: Task 'attempt_local408680812_0001_r_000000_0' done.
2025-05-20 13:58:10,197 INFO mapred.Task: Final Counters for attempt_local408680812_0001_r_000000_0: Counters: 30
File System Counters
FILE: Number of bytes read=7887
FILE: Number of bytes written=645606
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=75
HDFS: Number of bytes written=105
HDFS: Number of read operations=10
HDFS: Number of large read operations=0
HDFS: Number of write operations=3
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Combine input records=0
Combine output records=0
Reduce input groups=15
Reduce shuffle bytes=171
Reduce input records=15
Reduce output records=15
Spilled Records=15
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=0
Total committed heap usage (bytes)=526385152
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Output Format Counters
Bytes Written=105
2025-05-20 13:58:10,197 INFO mapred.LocalJobRunner: Finishing task: attempt_local408680812_0001_r_000000_0
2025-05-20 13:58:10,197 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-05-20 13:58:10,840 INFO mapreduce.Job: Job job_local408680812_0001 running in uber mode : false
2025-05-20 13:58:10,842 INFO mapreduce.Job: map 100% reduce 100%
2025-05-20 13:58:10,843 INFO mapreduce.Job: Job job_local408680812_0001 completed successfully


hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
Map input records=2
Map output records=15
Map output bytes=135
Map output materialized bytes=171
Input split bytes=101
Combine input records=0
Spilled Records=15
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=0
Total committed heap usage (bytes)=526385152
File Input Format Counters
Bytes Read=75
2025-05-20 13:58:10,020 INFO mapred.LocalJobRunner: Finishing task: attempt_local408680812_0001_m_000000_0
2025-05-20 13:58:10,020 INFO mapred.LocalJobRunner: map task executor complete.
2025-05-20 13:58:10,021 INFO mapred.LocalJobRunner: Waiting for reduce tasks
2025-05-20 13:58:10,022 INFO mapred.LocalJobRunner: Starting task: attempt_local408680812_0001_r_000000_0
2025-05-20 13:58:10,027 INFO output.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutp
utCommitterFactory
2025-05-20 13:58:10,027 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-05-20 13:58:10,027 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output d
irectory:false, ignore cleanup failures: false
2025-05-20 13:58:10,027 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2025-05-20 13:58:10,028 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffl
e@6622090a
2025-05-20 13:58:10,029 WARN impl.MetricsSystemImpl: JobTracker metrics system already initialized!
2025-05-20 13:58:10,037 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=5829453312, maxSingleShuffleLimit=14573633
28, mergeThreshold=3847439360, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-05-20 13:58:10,038 INFO reduce.EventFetcher: attempt_local408680812_0001_r_000000_0 Thread started: EventFetcher for f
etching Map Completion Events
2025-05-20 13:58:10,053 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local408680812_0001
_m_000000_0 decomp: 167 len: 171 to MEMORY
2025-05-20 13:58:10,054 INFO reduce.InMemoryMapOutput: Read 167 bytes from map-output for attempt_local408680812_0001_m_000
000_0
2025-05-20 13:58:10,055 INFO reduce.MergeManagerImpl: closeInMemoryFile -> map-output of size: 167, inMemoryMapOutputs.size
() -> 1, commitMemory -> 0, usedMemory ->167
2025-05-20 13:58:10,056 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning
2025-05-20 13:58:10,056 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 13:58:10,056 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outp
uts
2025-05-20 13:58:10,059 INFO mapred.Merger: Merging 1 sorted segments
2025-05-20 13:58:10,059 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 162 bytes
2025-05-20 13:58:10,060 INFO reduce.MergeManagerImpl: Merged 1 segments, 167 bytes to disk to satisfy reduce memory limit
2025-05-20 13:58:10,060 INFO reduce.MergeManagerImpl: Merging 1 files, 171 bytes from disk
2025-05-20 13:58:10,060 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2025-05-20 13:58:10,060 INFO mapred.Merger: Merging 1 sorted segments
2025-05-20 13:58:10,061 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 162 bytes
2025-05-20 13:58:10,061 INFO mapred.LocalJobRunner: 1 / 1 copied.


hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
2025-05-20 13:58:10,840 INFO mapreduce.Job: Job job_local408680812_0001 running in uber mode : false
2025-05-20 13:58:10,842 INFO mapreduce.Job: map 100% reduce 100%
2025-05-20 13:58:10,843 INFO mapreduce.Job: Job job_local408680812_0001 completed successfully
2025-05-20 13:58:10,854 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=15400
FILE: Number of bytes written=1291041
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=150
HDFS: Number of bytes written=105
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=2
Map output records=15
Map output bytes=135
Map output materialized bytes=171
Input split bytes=101
Combine input records=0
Combine output records=0
Reduce input groups=15
Reduce shuffle bytes=171
Reduce input records=15
Reduce output records=15
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=75
File Output Format Counters
Bytes Written=105
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /res/part-00000
cat: `/res/part-00000': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /res


hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~
Input split bytes=101
Combine input records=0
Combine output records=0
Reduce input groups=15
Reduce shuffle bytes=171
Reduce input records=15
Reduce output records=15
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=75
File Output Format Counters
Bytes Written=105
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /res/part-00000
cat: `/res/part-00000': No such file or directory
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /res
Found 2 items
-rw-r--r-- 1 hadoop supergroup 0 2025-05-20 13:58 /res/_SUCCESS
-rw-r--r-- 1 hadoop supergroup 105 2025-05-20 13:58 /res/part-r-00000
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ ^C
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /res/part-r-00000
college 1
in 1
bms 1
hii 1
i 1
inna 1
am 1
m 1
bhuvana 1
how 1
are 1
avyukth 1
of 1
you 1
engineeering 1
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$



Python - Hadoop Alphabetical order

Open hadoop ~~fold~~ folder:

Desktop-PC: ~/hadoop$ nano map.py

```
import sys
import re

for line in sys.stdin:
    words = re.findall(r'\b[a-z]+\b', line.lower())
    for word in words:
        print(f"{word}\t")
```

press ctrl x + Y + Enter

Desktop-PC: ~/hadoop$ nano red.py

```
import sys
import re

from collections import defaultdict

word_count = defaultdict(int)

for line in sys.stdin:
    word, count = line.strip().split('\t')
    word_count[word] += int(count)

sorted_words = sorted(word_count.items(),
                      key=lambda x: (-x[1], x[0]))[:10]

for word, count in sorted_words:
    print(f"{word}\t{count}")
```

Desktop-P: ~/hadoop$ nano probs.txt

Hello Hello my my in probs

Desktop-P: hadoop jar /home/hadoop/hadoop/share/hadoop/tools/lib/hadoop-streaming-3.3.4.jar -mapper map.py -reducer red.py -input file:///home/hadoop/hadoop/probs.txt -output file:///home/hadoop/hadoop/output1

output path

Lab 8: Scala

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

Code with Output:

```
bmscecse@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ nano pi.scala
bmscecse@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ scalac pi.scala
bmscecse@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ scala pi
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 5
7 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83
84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
```

Lab 9: Spark

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:

```
bmscecse@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ spark-shell
25/05/20 15:32:38 WARN Utils: Your hostname, bmscecse-HP-Elite-Tower-800-G9-Desktop-PC resolves to a loopback address: 127.0.1.1
; using 10.124.2.8 instead (on interface eno1)
25/05/20 15:32:38 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark/jars/spark-unsafe_2.12-3.0.3.jar) to con
structor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
25/05/20 15:32:38 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wh
ere applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
Spark context Web UI available at http://10.124.2.8:4040
Spark context available as 'sc' (master = local[*], app id = local-1747735361481).
Spark session available as 'spark'.
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 3.0.3
      /_/

Using Scala version 2.12.10 (OpenJDK 64-Bit Server VM, Java 11.0.26)
Type in expressions to have them evaluated.
Type :help for more information.

scala> val textFile = sc.textFile("/home/bmscecse/Desktop/sparkdata.txt")
textFile: org.apache.spark.rdd.RDD[String] = /home/bmscecse/Desktop/sparkdata.txt MapPartitionsRDD[1] at textFile at <console>:2
4

scala>

scala> val counts = textFile
counts: org.apache.spark.rdd.RDD[String] = /home/bmscecse/Desktop/sparkdata.txt MapPartitionsRDD[1] at textFile at <console>:24

scala>   .flatMap(line => line.split(" "))
res0: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at flatMap at <console>:26

scala>   .map(word => (word, 1))
scala> val data = sc.textFile("sparkdata.txt")
data: org.apache.spark.rdd.RDD[String] = sparkdata.txt MapPartitionsRDD[1] at textFile at <console>:25

scala> val splitdata = data.flatMap(line => line.split(" "))
splitdata: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at flatMap at <console>:26

scala> val mapdata = splitdata.map(word => (word, 1))
mapdata: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[3] at map at <console>:26

scala> val reducedata = mapdata.reduceByKey(_ + _)
reducedata: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[4] at reduceByKey at <console>:26

scala> reducedata.collect.foreach(println)
(,1)
(hello,2)
(world,1)
(spark,1)
```

```
scala> val textFile = sc.textFile("/home/bmscecse/Desktop/WC.txt")
textFile: org.apache.spark.rdd.RDD[String] = /home/bmscecse/Desktop/WC.txt MapPartitionsRDD[31] at textFile at <console>:31

scala> val words = textFile.flatMap(line => line.split(" "))
words: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[32] at flatMap at <console>:32

scala>

scala> val pairs = words.map(word => (word, 1))
pairs: org.apache.spark.rdd.RDD[(String, Int)] = MapPartitionsRDD[33] at map at <console>:32

scala>

scala> val counts = pairs.reduceByKey(_ + _)
counts: org.apache.spark.rdd.RDD[(String, Int)] = ShuffledRDD[34] at reduceByKey at <console>:32

scala> val countsArray = counts.collect()  // This is Array[(String, Int)]
countsArray: Array[(String, Int)] = Array(("",1), (hello,6), (world,1), (spark,1))

scala> val sorted = ListMap(countsArray.sortWith(_._2 > _._2): _*)
sorted: scala.collection.immutable.ListMap[String,Int] = ListMap(hello -> 6, "" -> 1, world -> 1, spark -> 1)

scala> for ((k, v) <- sorted) {
     |   if (v > 4) println(s"$k, $v")
     | }
hello, 6

scala>
```