

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

Sl. No.	Experiment Title	Page No.
1	MongoDB- CRUD Demonstration.	1
2	Perform the following DB operations using Cassandra. a) Create a keyspace by name Employee b) Create a column family by name Employee-Info with attributes Emp_Id Primary Key, Emp_Name, Designation, Date_of_Joining, Salary,Dept_Name c) Insert the values into the table in batch d) Update Employee name and Department of Emp-Id 121 e) Sort the details of Employee records based on salary f) Alter the schema of the table Employee_Info to add a column Projects which stores a set of Projects done by the corresponding Employee. g) Update the altered table to add project names. h) Create a TTL of 15 seconds to display the values of Employees.	6
3	Perform the following DB operations using Cassandra. a) Create a keyspace by name Library b) Create a column family by name Library-Info with attributes Stud_Id Primary Key, Counter_value of type Counter, Stud_Name, Book-Name, Book-Id, Date_of_issue c) Insert the values into the table in batch d) Display the details of the table created and increase the value of the counter e) Write a query to show that a student with id 112 has taken a book “BDA” 2 times. f) Export the created column to a csv file g) Import a given csv dataset from local file system into Cassandra column family	9
4	Execution of HDFS Commands for interaction with Hadoop Environment. (Minimum 10 commands to be executed)	11
5	Implement Wordcount program on Hadoop framework	14
6	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	23

	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.	28
8	Write a Scala program to print numbers from 1 to 100 using for loop.	35
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.	38
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).	39

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'  
  },  
]
```

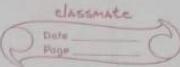
6. Find documents whose name begins with A

```
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: "Internetsurfing" })

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

db.Student.find({ StudName: "AryanDavid" })

db.Student.find({ StudName: "AryanDavid" })

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

db.Student.find({ Grade: { \$eq: "VII" } }).pretty()

db.Student.find({ StudName: /M/ }).pretty()

<pre>db.Student.count()</pre>	IX to set a particular field value to null <pre>db.Students.update({ _id: 3 }, { \$set: { Location: null } })</pre>
X Count the number of documents in Student Collections <pre>db.Students.count()</pre>	X Count the number of documents in Student Collections <pre>db.Students.count({ \$grade: "VII" })</pre>
XI Aggregate Function: <pre>db.Customers.aggregate([{ \$group: { _id: "FuntID", TotAcBal: { \$sum: "\$AcBal" } } }])</pre>	XI Count the number of documents in Student Collections with grade : VII <pre>db.Students.count({ \$grade: "VII" })</pre>
XII Create a collection Customers with fields custID, AcBal, <pre>db.Customers.aggregate([{ \$group: { _id: "FuntID", TotAcBal: { \$sum: "\$AcBal" } } }])</pre>	XII Aggregate Function: <pre>db.Customers.aggregate([{ \$match: { AcType: "S" } }, { \$group: { _id: "FuntID", TotAcBal: { \$sum: "\$AcBal" } } }])</pre> <p style="text-align: right;">888</p> <p style="text-align: right;">113125</p>
VII Remove the field in an existing Document <pre>db.Student.find({ _id: 1 }, { StudName: 1, Grade: 1, _id: 0 })</pre>	VII Remove the field in an existing Document <pre>db.Student.update({ _id: 4 }, { \$set: { Location: "Network" } })</pre>
VI Add a new field to existing Document <pre>db.Students.update({ _id: 4 }, { \$set: { Location: "Network" } })</pre>	VI Add a new field to existing Document <pre>db.Students.update({ _id: 4 }, { \$set: { Location: "Network" } })</pre>
V Save Method: <pre>db.Students.save({ StudName: "Vamsi", Grade: "VI" })</pre>	V Save Method: <pre>db.Students.save({ StudName: "Vamsi", Grade: "VI" })</pre>
IV Export data to a CSV file <pre>mongoexport --host localhost --db Student --collection airlines --type csv -out /home/hduser/Desktop/output.txt --fields "Year", "Quarter"</pre>	IV Import data from a CSV file <pre>mongorestore --db Student --collection airlines --type csv -h localhost -f /home/hduser/Desktop/airline.csv</pre>

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

✓ 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: { $last: "$Cust_id" },
    min_balance: { $min: "$Acc_Bal" },
    max_balance: { $max: "$Acc_Bal" }
  }}])
```

Output: [{ _id: 3, min_balance: 1300, max_balance: 2000 }, { _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 }]

IV

✓ Create the Products collection

use `eCommerce`:

`db.createCollection("Products")`

✓ db.Products.insertMany([

- `{ product_id: "prod123",`
- `name: "Smartphone",`
- `category: "Electronics",`
- `price: 499,`
- `quantity: 50 }`

])

1. product_id: "prod124",
 name: "Laptop",
 category: "Electronics",
 price: 899,
 quantity: 30
]
 2. product_id: "prod125",
 name: "Headphones",
 category: "Electronics",
 price: 799,
 quantity: 100
]
 :
 3.
 2. Create Carts Collection:
 db.createCollection("carts");
 db.carts.insertMany([
 { user_id: "xyzghi",
 products: [
 { product_id: "prod123", quantity: 2 },
 { product_id: "prod124", quantity: 1 }
]
 },
 { user_id: "456zyx",
 products: [
 { product_id: "prod124", quantity: 1 },
 { product_id: "prod125", quantity: 3 }
]
 },
 { user_id: "123abc",
 products: [
 { product_id: "prod123", quantity: 2 }
]
 }
])

3. Create the Orders Collection
 db.createCollection("orders")
 db.orders.insertMany([
 { user_id: "123abc",
 products: [
 { product_id: "prod123", quantity: 2, price: 899 },
 { product_id: "prod125", quantity: 1, price: 799 }
],
 total_price: 1698,
 status: "Placed",
 order_date: ISODate("2023-3-11T15:00:00Z")
 },
 { user_id: "456zyx",
 products: [
 { product_id: "prod124", quantity: 1, price: 799 }
],
 total_price: 799,
 status: "Placed",
 order_date: ISODate("2023-3-11T15:00:00Z")
 },
 { user_id: "xyzghi",
 products: [
 { product_id: "prod123", quantity: 2, price: 899 }
],
 total_price: 1798,
 status: "Placed",
 order_date: ISODate("2023-3-11T15:00:00Z")
 }
])

4. Retrieved:
 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": { \$le: 100 } });

6. Retrieve Orders Placed by a user (User ID with "123abc")
 db.Orders.find({ user_id: "123abc" });

7. Retrieve 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 }] } } } }
])

8/12/23
 11/12/23

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

Create Keyspace:

```
CREATE KEYSPACE Students WITH REPLICATION = 'SimpleStrategy' , 'replication_factor':3;
```

DESCRIBE KEYSPACES:

```
SELECT * FROM system.schema.keyspaces;
```

USE Students;

```
CREATE TABLE Students_Info(Roll_No int PRIMARY KEY, StudName text, DateOfJoining timestamp, last_exam_percent double);
```

DESCRIBE TABLES;

```
DESCRIBE TABLE <table> Students_Info;
```

Insert:

```
BEGIN BATCH
INSERT INTO Students_Info(Roll_No, StudName, DateOfJoining, last_exam_percent)
VALUES (1, 'Asha', '2012-03-12', 79.9)
INSERT INTO Students_Info(Roll_No, StudName, DateOfJoining, last_exam_percent) VALUES
VALUES (2, 'Samith', '2012-03-12', 90.9)
INSERT INTO Students_Info(Roll_No, StudName, DateOfJoining, last_exam_percent)
VALUES (3, 'Smitha', '2012-03-12', 89.9)
```

APPLY BATCH;

(1) SELECT * FROM Students_Info;

```
SELECT * FROM Students_Info WHERE Roll_No IN (1,2,3),
```

select * from students_info where Studname = 'Asha';

CREATE INDEX ON Students_Info (StudName);

select Roll_No, StudName from students_info LIMIT 2;

select Roll_No as "USN" from Students_Info;

UPDATE Students_Info SET StudName = 'David Shein' WHERE RollNo = 2;

(1) Output :

roll-no	dateofjoining	last_exam_percent	studname
1	2012-03-11 18:30	79.9	Asha

(2) roll-no | dateOfjoining | last_exam_percent | Studname

roll-no	dateOfjoining	last_exam_percent	Studname
2	2012-03-11	90.9	Tanu

(3) rollno | dateOfjoining | last_exam_percent | studname

rollno	dateOfjoining	last_exam_percent	studname
3	2012-03-12 18:30	89.9	Asha

UPDATE Students_Info SET roll_no=6 where rollno=3;

DELETE

Lab 4: Cassandra

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

Code with Output:

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

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

```
hadoop@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ touch test.txt
hadoop@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ nano text.txt
hadoop@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -put ./text.txt /Lab05/text.txt
hadoop@bmscsecse-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@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -cat /Lab05/text.txt
Hello
How are you?
```

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

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

```
hadoop@bmscsecse-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@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~/Desktop$ hdfs dfs -cp /test_Lab05/ /Lab05
hadoop@bmscsecse-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@bmscsecse-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_val counter,
        primary key ((stud_id, book_id), stud_name,
        bookname, 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 + 1 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
	2	S1	B2	2024-01-01	0

→ To indicate that student 112 has taken ADA 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 = "ADA"
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
112	11	S112	ADA	2024-01-02	2

→ Import and Export to CSV

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

Page

15/4/24

cat sample.csv	
1, 1, S1, B1, 2024-01-01, 1	1.
1, 2, S1, B2, 2024-01-01, 1	
112, 11, S112, BDA, 2024-01-02, 2	
]	
:	
> drop truncate library_info;	
> copy library_info from except 'sample.csv';	2.
> select * from table where stud_id = 112 allow filtering;	3.
1 Stud-id book-id Stud-name Book-name Issue-date Contdental	
112 11 S112 BDA 2024-01-02 2	
15/4	4

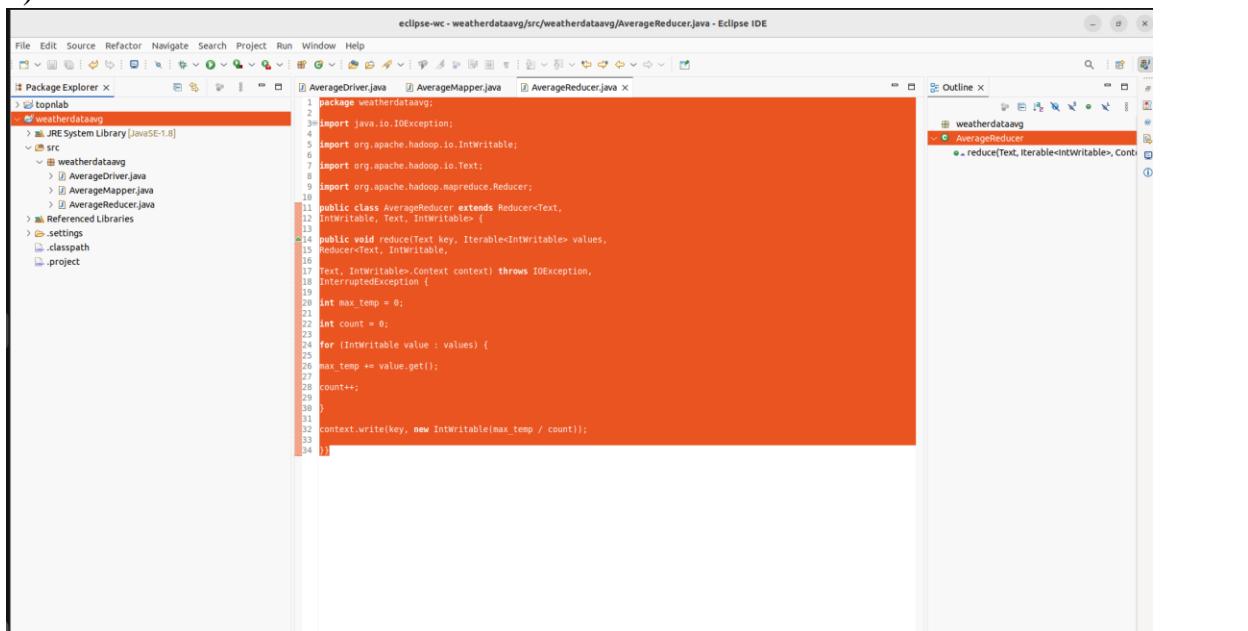
Lab 5: Hadoop

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

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

Code with Output:

a)



The screenshot shows the Eclipse IDE interface with the project 'weatherdataavg' selected in the Package Explorer. The AverageReducer.java file is open in the editor. The code implements a Reducer that calculates the average temperature for each year. It iterates over the input values, summing them up and dividing by the count to find the average. The code is annotated with comments explaining the logic.

```
1 package weatherdataavg;
2
3 import java.io.IOException;
4 import org.apache.hadoop.io.IntWritable;
5 import org.apache.hadoop.io.Text;
6 import org.apache.hadoop.mapreduce.Reducer;
7
8 public class AverageReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
9
10    public void reduce(Text key, Iterable<IntWritable> values,
11                      Reducer<Text, IntWritable, Text, IntWritable>.Context context) throws IOException,
12                                          InterruptedException {
13
14        int max_temp = 0;
15
16        int count = 0;
17
18        for (IntWritable value : values) {
19            max_temp += value.get();
20
21            count++;
22
23        }
24
25        context.write(key, new IntWritable(max_temp / count));
26
27    }
28
29 }
30
31 }
```




The terminal window shows the execution of the Hadoop command to start the cluster and run the AverageDriver job. The output includes logs for the NameNode, DataNodes, and Secondary NameNode, followed by the results of the wordcount operation on the 'wordcount' file.

```
hadoop@bmscse-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 recommended production deployment configuration.
Hadoop is now running on port 54321.
Starting namenodes [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 [bmscse-HP-Elite-Tower-800-G9-Desktop-PC]
bmscse-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 job 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@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mkdir /kln
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -copyFromLocal /home/hadoop/Desktop/1901 /kln/1901
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/weatherdataavg.jar weatherdataavg.AverageDriver /kln/1901/rem
Please enter the input and output parameters
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ ^C
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -rm -r /remout
rm: '/remout': No such file or directory
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /
Found 9 items
drwxr-xr-x - hadoop supergroup 0 2025-04-15 15:11 /CSE
drwxr-xr-x - hadoop supergroup 0 2025-04-15 15:11 /LLI
drwxr-xr-x - hadoop supergroup 0 2024-05-14 15:39 /abc
drwxr-xr-x - hadoop supergroup 0 2025-05-20 14:30 /kln
drwxr-xr-x - hadoop supergroup 0 2025-05-20 13:52 /mno
drwxr-xr-x - hadoop supergroup 0 2025-05-20 13:58 /res
drwxr-xr-x - hadoop supergroup 0 2025-04-29 15:38 /rgs
drwxr-xr-x - hadoop supergroup 2024-05-21 15:37 /wordcount
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /kln/
Found 1 item
-rw-r--r-- 1 hadoop supergroup 888190 2025-05-20 14:30 /kln/1901
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /kln/1901
```

```

hadoop@bmseccse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/weatherdataavg.jar weatherdataavg.AverageDriver /kml/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 14:34:21,230 INFO input.FileInputFormat: Input split number of splits=1
2025-05-20 14:34:21,334 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local261815800_0001
2025-05-20 14:34:21,400 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.FileOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:34:21,406 INFO output.FileOutputCommitterFactory: FileOutputCommitter 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: org.apache.hadoop.mapred.lib.output.FileOutputCommitter
2025-05-20 14:34:21,440 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.FileOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:34:21,460 INFO output.FileOutputCommitter: FileOutputCommitter 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 14:34:21,460 INFO mapred.LocalJobRunner: org.apache.hadoop.mapred.lib.output.FileOutputCommitter
2025-05-20 14:34:21,460 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-05-20 14:34:21,472 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/kml/1901+0:888190
2025-05-20 14:34:21,511 INFO mapred.MapTask: (EQUATOR) 0 kv1 26214396(104857584)
2025-05-20 14:34:21,511 INFO mapred.MapTask: mapred.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 14:34:21,511 INFO mapred.MapTask: mapred.mapoutput.collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-20 14:34:21,582 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 14:34:21,595 INFO mapred.MapTask: 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 small 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 records=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_TYPE=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: reducing task executor complete.
2025-05-20 14:34:22,001 INFO mapred.LocalJobRunner: Job attempt_local261815800_0001 running in uber mode : false
2025-05-20 14:34:22,405 INFO mapred.Job: map 100% reduce 100%
2025-05-20 14:34:22,407 INFO mapreduce.Job: Job job_local261815800_0001 completed successfully
2025-05-20 14:34:22,416 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=153112
FILE: Number of bytes written=1500206
FILE: Number of small 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=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
Combined Map/ReduceRecords=6564
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
  Bytes Written=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: output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:34:21,612 INFO FileOutputCommitter: FileOutputCommitter Algorithm version is 2
2025-05-20 14:34:21,612 INFO FileOutputCommitter: Using ResourceCalculatorProcessTree
2025-05-20 14:34:21,612 INFO mapred.Task: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@cd4d167
2025-05-20 14:34:21,614 WARN Iml.MetricsSystemImpl: Jobtracker metrics system already initialized!
2025-05-20 14:34:21,622 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=5829453312, maxSingleShuffleLimit=1457363328, mergeThreshold=3847439360, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-05-20 14:34:21,623 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=5829453312, maxSingleShuffleLimit=1457363328, mergeThreshold=3847439360, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-05-20 14:34:21,635 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map attempt_local261815800_0001_m_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 decom: 72206 len: 72210 to MEMORY
2025-05-20 14:34:21,637 INFO reduce.MergeManagerImpl: Closing temporary file -> 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.Task: Task is interrupted.. Returning
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 14:34:21,641 INFO reduce.MergeManagerImpl: Merging 1 segments, 72206 bytes to disk to satisfy reduce memory limit
2025-05-20 14:34:21,643 INFO reduce.MergeManagerImpl: Merging 0 segments, 72206 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.output.dir deprecated. Instead, use mapreduce.job.skipreduces
2025-05-20 14:34:21,679 INFO mapred.Task: attempt_local261815800_0001_r_000000_0 is done. And is in the process of committing
2025-05-20 14:34:21,680 INFO mapred.Task: attempt_local261815800_0001_r_000000_0 is committed.
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=1484882
  FILE: Number of bytes written=706208
  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)=1266679888
Shuffle Errors
  BAD_IDs=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@bmseccse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /rem/part-00000
cat: '/rem/part-00000': No such file or directory
hadoop@bmseccse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /rem/part-r-00000
1901      46
hadoop@bmseccse-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)

The screenshot shows the Eclipse IDE interface with the following details:

- Eclipse IDE Title Bar:** eclipse-wc - weatherdatameanmax/src/weatherdatameanmax/MeanMaxReducer.java - Eclipse IDE
- Package Explorer:** Shows the project structure with packages: topnlab, weatherdataavg, and weatherdatameanmax. Under weatherdatameanmax, there are sub-packages: JRE System Library [JavaSE-1.8], src, and Referenced Libraries.
- MeanMaxReducer.java Content:** The code implements a Reducer for Hadoop. It initializes variables for maxTemp, totalTemp, count, and days. It iterates through a list of IntWritable values, updating maxTemp if the current value is greater. After every three iterations, it calculates the average and writes it to the context. If there are fewer than three values, it writes zero.
- Outline View:** Shows the class structure with the reduce method highlighted.
- Terminal Window:** Displays the output of the "start-all.sh" command on a Hadoop cluster. The log includes messages about starting namenodes, datanodes, secondarynamenode, resourcemanager, and nodemanagers. It also lists various Hadoop commands like archive, checknative, config, debug, help, hosts, loglevel, workers, and jar.

```
hadoop@bmsece-HP-Elite-Tower-800-G9-Desktop-PC: ~

jnpipath applications, not this command.
prints the java.library.path
kdlag Diagnose Kerberos Problems
kerbname show auth_to_local principal conversion
key manage keys via the KeyProvider
rumenfilter scale a rumen input trace
rumentrace convert logs into a rumen trace
3sguard 53 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@bmsece-HP-Elite-Tower-800-G9-Desktop-PC: ~$ hadoop fs -mkdir /omn
hadoop@bmsece-HP-Elite-Tower-800-G9-Desktop-PC: ~$ hadoop fs -copyFromLocal /home/hadoop/Desktop/1901 /omn/1901
hadoop@bmsece-HP-Elite-Tower-800-G9-Desktop-PC: ~$ hadoop jar /home/hadoop/Desktop/weatherdatameanmax.jar weatherdatameanmax.MeanMaxDriver /omn/1901 /ren
2025-05-20 14:53:15,512 INFO [impl.MetricsConfig] Loaded properties from hadoop-metrics2.properties
2025-05-20 14:53:15,615 INFO [impl.MetricsConfig] Schema version: 1
2025-05-20 14:53:15,615 INFO [impl.MetricsConfig] Snapshot period at 10 second(s).
2025-05-20 14:53:15,615 INFO [impl.MetricsConfig] Metrics system initialized
2025-05-20 14:53:15,679 WARN mapred.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 14:53:15,823 INFO [mapreduce.JobSubmitter] Submitting tokens for job: job_local2143084439_0001
2025-05-20 14:53:15,823 INFO [mapreduce.JobSubmitter] Executing with tokens: []
2025-05-20 14:53:15,899 INFO [mapreduce.Job] The url to track the job: http://localhost:8080/
2025-05-20 14:53:15,901 INFO [mapreduce.Job] Job ID: job_local2143084439_0001
2025-05-20 14:53:15,901 INFO [mapred.LocalJobRunner] OutputCommitter set in config null
2025-05-20 14:53:15,905 INFO [output.PathOutputCommitterFactory] No output committer factory defined, defaulting to PathOutputCommitterFactory
2025-05-20 14:53:15,905 INFO [output.PathOutputCommitter] File Output Committer Algorithm version is 2
2025-05-20 14:53:15,905 INFO [output.PathOutputCommitter] FileOutputCommitter: skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:53:15,905 INFO [output.PathOutputCommitter] 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 14:53:15,981 INFO [mapred.LocalJobRunner] Starting task attempt_local2143084439_0001_m_000000_0
2025-05-20 14:53:15,981 INFO [mapred.Task] Task attempt_local2143084439_0001_m_000000_0 starting in map 0
2025-05-20 14:53:15,996 INFO [output.FileOutputCommitter] File 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 14:53:16,004 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 kvlt 26214396(104857584)
2025-05-20 14:53:16,044 INFO [mapred.MapTask] mapred.Task: mapreduce.task.io sort.mb: 100
2025-05-20 14:53:16,044 INFO [mapred.MapTask] mapred.Task: sortLimit = 83886080
2025-05-20 14:53:16,044 INFO [mapred.MapTask] mapred.Task: bufSize = 0; bufEnd = 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 14:53:16,120 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 14:53:16,133 INFO [mapred.LocalJobRunner] minid -> localJobRunner.info

hadoop@bmsece-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, lgnore 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@64ffe0
2025-05-20 14:53:16,147 WARN [impl.MetricsSystemImpl] JobTracker metrics system already initialized!
2025-05-20 14:53:16,153 INFO [reduce.MergeManagerImpl] MergeManager: memoryLmt=5829453312, maxSingleShuffleLmt=145730328, mergeThreshold=3847439360, toSortFactor=10, memToMemMergeOutputsThreshold=10
2025-05-20 14:53:16,170 INFO [reduce.LocalJobContext] localfetcher@1 about to shuffle output of map attempt_local2143084439_0001_m_000000_0 Thread started: EventFetcher for fetching Map Completion Events
2025-05-20 14:53:16,170 INFO [reduce.LocalJobContext] localfetcher@1 about to shuffle output of map attempt_local2143084439_0001_m_000000_0 memToMemMergeOutputsThreshold=10
2025-05-20 14:53:16,172 INFO [reduce.MergeManagerImpl] Closing temporary file -> 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 14:53:16,174 INFO [mapred.Merger] Merging 1 sorted segments
2025-05-20 14:53:16,174 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] Merging 1 segments, 59078 bytes to disk to reduce memory limit
2025-05-20 14:53:16,180 INFO [reduce.MergeManagerImpl] Reducing memory usage by 1 file, 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 ls deprecated. Instead, use mapreduce.job.skippedrecords
2025-05-20 14:53:16,212 INFO [mapred.Task] Task attempt_local2143084439_0001_r_000000_0 is done. And is in the process of committing
2025-05-20 14:53:16,212 INFO [mapred.Task] Task attempt_local2143084439_0001_r_000000_0 is done.
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=1913
  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 output records=6564
  Reduce output records=12
  Spilled Records=5564
  Shuffled Maps =1
  Failed Shuffles=0
  Merged Map outputs=1
  GC time elapsed (ns)=0
  Total committed heap usage (bytes)=526385152
Shuffle
  END_OF_SHUFFLE=0
  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_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 14:53:16,007 INFO mapred.MapTask: Processing split: hdfs://localhost:9000/omn/1901:0+888190
2025-05-20 14:53:16,044 INFO mapred.MapTask: EQUATOR: 0 kv1 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 14:53:16,044 INFO mapred.MapTask: MapOutputCollector: org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-20 14:53:16,119 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 spilt 0
2025-05-20 14:53:16,135 INFO mapred.LocalJobRunner: 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=0
2025-05-20 14:53:16,139 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 output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
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
  Failed Maps =1
  Shuffled Maps =0
  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_PARTITION=0
  File Output Format Counters
  Bytes Written=81
2025-05-20 14:53:16,299 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 14:53:16,900 INFO mapreduce.Job: Counters: 30
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=1776988
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=0

```

```

hadoop@bmscse-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=177088
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
  Reducer input records=0
  Reduce shuffle bytes=59882
  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)=8
  Total committed heap usage (bytes)=1052770304
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAGIC=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=888190
  File Output Format Counters
    Bytes Written=81
hadoop@bmscse-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@bmscse-HP-Elite-Tower-800-G9-Desktop-PC: $ hadoop fs -cat /ren/part-r-00000
01 13
02 -66
03 -15
04 43
05 100
06 186
07 219
08 198
09 141
10 100
11 1
12 -61
hadoop@bmscse-HP-Elite-Tower-800-G9-Desktop-PC: $
```

15/4/24 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
 drwxr-xr-x - hadoop supergroup 0 1144-MM-20
 2. MH .MM / bda-hadoop
 4. \$ hdfs dfs -put /home/hduser/Desktop/bda-local.txt
 bda-hadoop/file.txt
 5. \$ hdfs dfs -cat /bda-hadoop/file.txt
 Output : Hello Prakharjan
 → Copy from local
 6. \$ hdfs dfs -copyFromLocal /home/hduser/Desktop/bda-local
 bda-hadoop/file.txt
 7. \$ hdfs dfs -cat /bda-hadoop/file_local.txt
 Output : Hello prae Prakharjan

8. \$ hdfs dfs -get -get /bda-hadoop/file.txt /home/hadoop
Downloaded file.txt

9. \$ hdfs dfs -getmerge /bda-hadoop/file.txt /bda-hadoop/
file_cp-local.txt /home/hadoop/Desktop/

10. \$ hadoop fs -getfacl /bda-hadoop/
Output:
#file: /bda-hadoop
owner: hduven
group: supengroup
user::rwx
group::r-x
other::r-x

11. \$ hdfs dfs -copyToLocal /bda-hadoop/file.txt /home/hadoop/Desktop

12. \$ hadoop fs -mr /bda-hadoop/abc

13. \$ hadoop fs -ls /abc

Output: Found 1 items
drwxr-xr-x - hduven supengroup 0 4444-10-
HH:MM /abc/bda-hadoop

14. \$ hadoop fs -cp /hello/hadoop_lab
① 24/10/25

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
    year, temperature = line.split('\t')
    if current_year == None:
        current_year = year
        temp_min = temperature
        temp_max = temperature
        temp_sum = float(temperature)
        temp_count += 1
    else:
        temp_sum += float(temperature)
        temp_count += 1
        temp_min = min(temp_min, float(temperature))
        temp_max = max(temp_max, float(temperature))

if current_year is not None:
    avg_temp = temp_sum / temp_count if temp_count > 0 else 0
    print(f'{current_year}\tMin:{temp_min}\tAvg:{avg_temp}\tMax:{temp_max}')
    print(f'Avg:{avg_temp}')

```

In shell

Desktop-Pc : \$ hadoop jar /home/hadoop/hadoop-streaming-3.2.0.jar
 hadoop/tools/lib/hadoop-streaming-3.2.0.jar
 -mapper map.py
 -reducer red.py
 -input /lab0/in.txt
 -output /lab0/output3

....:\$ hdfs dfs -ls /lab0/output3
 "(for checking)"
 \$ hdfs dfs -cat /lab0/output3/part-00
 print output

Lab 6: Hadoop

Question: Implement Wordcount program on Hadoop framework

Code with Output:

The screenshot shows the Eclipse IDE interface with the following details:

- Project Explorer:** Shows the `word_count` project with a `src` folder containing `WCDriver.java`, `WCMapper.java`, and `WCReducer.java`.
- Code Editor:** Displays the `WCDriver.java` file with the following code:import java.io.IOException;
import org.apache.hadoop.conf.Configured;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.FileInputFormat;
import org.apache.hadoop.mapred.FileOutputFormat;
import org.apache.hadoop.mapred.JobClient;
import org.apache.hadoop.mapred.JobConf;
import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;
public class WCDriver extends Configured implements Tool {
 public int run(String[] args) throws IOException {
 if (args.length < 2) {
 System.out.println("Please provide input and output paths");
 return -1;
 }
 JobConf conf = new JobConf(WCDriver.class);
 conf.setJobName("WordCount");
 conf.setJarByClass(WCDriver.class); // Ensures job runs from correct JAR
 FileInputFormat.setInputPaths(conf, new Path(args[0]));
 FileOutputFormat.setOutputPath(conf, new Path(args[1]));
 conf.setMapperClass(WCMapper.class);
 conf.setReducerClass(WCReducer.class);
 conf.setMapOutputKeyClass(Text.class);
 conf.setMapOutputValueClass(IntWritable.class);
 }
}
- Problems View:** Shows a message about Java 24 being released and offers to install Java 24 Support.
- Console View:** Displays the terminal output of the Hadoop WordCount application, showing the process of creating a directory, copying a sample file, running the jar, and displaying the results.

```
hadoop@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mkdir /rsg
hadoop@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -copyFromLocal /home/hadoop/Desktop/sample.txt /rsg/sample.txt
hadoop@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/wordcount.jar WCDriver /rsg/sample.txt /result
2025-05-06 15:05:01,260 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
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 interface 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
```

```

hadoop@bmscse-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 interface 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 kvt 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 15:05:01,751 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@bmscse-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

```

```

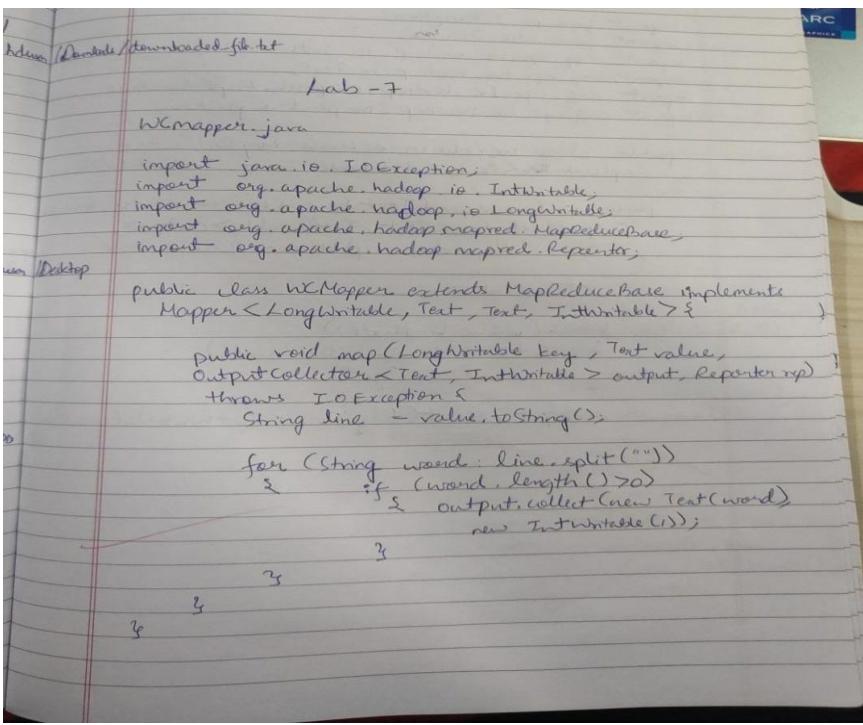
hadoop@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC: ~
FILE: Number of write operations=0
HDFS: Number of bytes read=89
HDFS: Number of bytes written=0
HDFS: Number of read operations=5
HDFS: Number of large read operations=0
HDFS: Number of write operations=1
HDFS: Number of bytes read erasure-coded=0
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, mergeThreshold=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 Completion 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 decomp: 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, committedMemory -> 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@bmscsece-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@bmscsece-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@bmscsece-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@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC: ~$ []

```



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;
import org.apache.hadoop.mapreduce.*;

public class WCReducer extends MapReduceBase
    implements <Text, IntWritable, Text, IntWritable> {
    public void reduce(Text key, Iterator<IntWritable> value,
                      OutputCollector<Text, IntWritable> output,
                      Reporter reporter) 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.Configuration;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.mapred.JobConf;
public class WCDriver extends Configuration implements
    Tool {
    public int run(String args[]) throws IOException {
        if (args.length < 2) {
            System.out.println("Please give
                valid input");
            return -1;
        }
    }
}
```

```
JobConf conf = new JobConf(WCDriver.class),
FileInputFormat.setInputPaths(conf, new Path(args[0]));
FileOutputFormat.setOutputPath(conf, new Path(args[1]));
conf.setMapperClass(WCMapper.class),
conf.setReducerClass(WCReducer.class),
conf.setMapOutputValueClass(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);
}
```

(10/29/14) System.out.println(exitCode);
3

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:~$ 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 node 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.txt
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@bmscsecse-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@bmscsecse-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.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@bmscsecse-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$1.run(URLClassLoader.java:452)
at java.base/java.security.AccessController.doPrivileged(Native Method)
at java.base/java.net.URLClassLoader.findClass(URLClassLoader.java:451)
at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:594)
at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:527)
at java.base/java.lang.Class.forName0(Native Method)
at java.base/java.lang.Class.forName(Class.java:398)
at org.apache.hadoop.util.RunJar.run(RunJar.java:321)
at org.apache.hadoop.util.RunJar.main(RunJar.java:241)
hadoop@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ 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@bmscse-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 directory:false, ignore cleanup failures: false
2025-05-20 13:58:09,842 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
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 FileOutputCommitterFactory
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 directory: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 = 57/6553600
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 committing
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@bmscsece-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 committing
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 hdfs://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@bmscsece-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 committing
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 hdfs://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@bmscscse-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 FileOutputCommitterFactory
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 directory: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.Shuffle@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=14573633328, 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 fetching 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_000000_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-outputs
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@bmscsecse-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@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /res/part-00000
cat: '/res/part-00000': No such file or directory
hadoop@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -ls /res
```

```
hadoop@bmscsecse-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@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /res/part-00000
cat: '/res/part-00000': No such file or directory
hadoop@bmscsecse-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@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ ^C
hadoop@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -cat /res/part-r-00000
college 1
in 1
bms 1
hi 1
i 1
inna 1
am 1
m 1
bhuvana 1
how 1
are 1
avyukth 1
of 1
you 1
engineering 1
hadoop@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~$
```

classmate
Date _____
Page _____

Python - Hadoop Alphabetical order

Open hadoop folder:

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

```
input sys
import re

for line in sys.stdin:
    words = re.findall(r'\b[a-zA-Z]+\b', line.lower())
    for word in words:
        print(f'{word}\t1')
```

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}')
```

classmate
Date _____
Page _____

Desktop-P: - / hadoop \$ nano probst.txt

Hello Hello my my in probst

JP
CD 30/10/15

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/probst.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
```

The screenshot shows a terminal window titled "bmscecse@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC: ~". Inside the terminal, the code for "pi.scala" is displayed in a nano text editor. The code defines an object "pi" with a main method that prints numbers from 1 to 100. Below the code, the terminal shows the output of running the Scala interpreter on the file. At the bottom of the terminal window, there is a menu bar with various keyboard shortcuts for nano, such as "Help", "Exit", "Read File", "Replace", "Cut", "Paste", "Execute", "Justify", "Location", and "Go To Line".

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

[Read 7 lines]

^G Help ^O Write Out ^W Where Is ^K Cut ^T Execute ^C Location
^X Exit ^R Read File ^\ Replace ^U Paste ^J Justify ^/ Go To Line

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 occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark/jars/spark-unsafe_2.12-3.0.3.jar) to constructor 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 where 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

    /_\_/\_ _\_\_/\_/\_/\_
   / \ \ - \ \ - / \ - / \ /
  /__\ / . \ \_/_ / \ / \ \_ \
 /_/

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>

```

