

VISVESVARAYA TECHNOLOGICAL UNIVERSITY

“JnanaSangama”, Belgaum -590014, Karnataka.



LAB REPORT

on

Big Data Analytics(23CS6PCBDA)

Submitted by

Raghavendra N (1BM22CS213)

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 **Raghavendra N (1BM22CS037)**, who is bonafide student of **B. M. S. College of Engineering**. It is in partial fulfillment for the award of **Bachelor of Engineering in Computer Science and Engineering** of the Visvesvaraya Technological University, Belgaum during the year 2025. The Lab report has been approved as it satisfies the academic requirements in respect of a **Big Data Analytics(23CS6PCBDA)** work prescribed for the said degree.

Pradeep Sadanand
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	1-4
2	MongoDB	5-8
3	Neo4j	9-12
4	Cassandra	12-20
5	Cassandra	21-29
6	Hadoop	30-37
7	Hadoop	38-43
8	Scala and Spark	44-48

https://github.com/RaghavendraN-cs213/BDA_LAB

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-ws5rct-shard-0 [primary] mydb> db.createCollection('Customers')
{ ok: 1 }

Atlas atlas-ws5rct-shard-0 [primary] mydb> db.createCollection('Student')
{ ok: 1 }

Atlas atlas-ws5rct-shard-0 [primary] mydb> db.createCollection('Student')
{ ok: 1 }

Atlas atlas-ws5rct-shard-0 [primary] test> use mydb
switched to db mydb

Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.deleteMany({Grade:'VII'})
{ acknowledged: true, deletedCount: 3 }

Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.deleteOne({StudName:'JacobAdam'})
{ acknowledged: true, deletedCount: 0 }

Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.drop()
true

Atlas atlas-ws5rct-shard-0 [primary] mydb> db.dropDatabase()
{ ok: 1, dropped: 'mydb' }

Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.remove({StudName:'JacobAdam'})
DeprecationWarning: Collection.remove() is deprecated. Use deleteOne, deleteMany, findOneAndDelete, or bulkWrite.
{ acknowledged: true, deletedCount: 0 }
```

```
[Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Customers.find()]
[  {
      _id: ObjectId('67c6c71f812483cc27dd4a64'),
      cust_id: 1,
      balance: 200,
      type: 'S'
    },
    {
      _id: ObjectId('67c6c739812483cc27dd4a65'),
      cust_id: 1,
      balance: 1000,
      type: 'Z'
    },
    {
      _id: ObjectId('67c6c74d812483cc27dd4a66'),
      cust_id: 2,
      balance: 100,
      type: 'Z'
    },
    {
      _id: ObjectId('67c6c75e812483cc27dd4a67'),
      cust_id: 2,
      balance: 1000,
      type: 'C'
    },
    {
      _id: ObjectId('67c6c76e812483cc27dd4a68'),
      cust_id: 2,
      balance: 500,
      type: 'C'
    },
    {
      _id: ObjectId('67c6c781812483cc27dd4a69'),
      cust_id: 2,
      balance: 50,
      type: 'S'
    },
    {
      _id: ObjectId('67c6c795812483cc27dd4a6a'),
      cust_id: 3,
      balance: 500,
      type: 'Z'
    }
]
```

Google Classroom

```
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.find()
[
  {
    _id: ObjectId('67c6c3c3812483cc27dd4a5d'),
    RollNo: 1,
    Age: 21,
    Cont: 9876,
    email: 'antara.de9@gmail.com'
  },
  {
    _id: ObjectId('67c6c3d8812483cc27dd4a5e'),
    RollNo: 1,
    Age: 21,
    Cont: 9876,
    email: 'antara.de9@gmail.com'
  },
  {
    _id: ObjectId('67c6c458812483cc27dd4a5f'),
    RollNo: 2,
    Age: 22,
    Cont: 9976,
    email: 'anushka.de@gmail.com'
  },
  {
    _id: ObjectId('67c6c47f812483cc27dd4a60'),
    RollNo: 3,
    Age: 21,
    Cont: 5576,
    email: 'anubhav.de@gmail.com'
  },
  {
    _id: ObjectId('67c6c4a2812483cc27dd4a61'),
    RollNo: 4,
    Age: 20,
    Cont: 4476,
    email: 'pani.de9@gmail.com'
  },
  {
    _id: ObjectId('67c6c4db812483cc27dd4a62'),
    RollNo: 10,
    Age: 23,
    Cont: 2276,
    email: 'rekha.de9@gmail.com'
  }
]
```

```
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.find()
[[
  {
    _id: ObjectId('67c6c3c3812483cc27dd4a5d'),
    RollNo: 1,
    Age: 21,
    Cont: 9876,
    email: 'antara.de9@gmail.com'
  },
  {
    _id: ObjectId('67c6c3d8812483cc27dd4a5e'),
    RollNo: 1,
    Age: 21,
    Cont: 9876,
    email: 'antara.de9@gmail.com'
  },
  {
    _id: ObjectId('67c6c458812483cc27dd4a5f'),
    RollNo: 2,
    Age: 22,
    Cont: 9976,
    email: 'anushka.de@gmail.com'
  },
  {
    _id: ObjectId('67c6c47f812483cc27dd4a60'),
    RollNo: 3,
    Age: 21,
    Cont: 5576,
    email: 'anubhav.de@gmail.com'
  },
  {
    _id: ObjectId('67c6c4a2812483cc27dd4a61'),
    RollNo: 4,
    Age: 20,
    Cont: 4476,
    email: 'pani.de9@gmail.com'
  },
  {
    _id: ObjectId('67c6c4db812483cc27dd4a62'),
    RollNo: 10,
    Age: 23,
    Cont: 2276,
    email: 'abhinav@gmail.com'
  },
]]
```

```
{  
  _id: ObjectId('67c6c616812483cc27dd4a63'),  
  RollNo: 11,  
  Age: 22,  
  Name: 'FEM',  
  cont: 2276,  
  email: 'rea.de9@gmail.com'  
},  
{  
  _id: 1,  
  StudName: 'Michelle Jacintha',  
  Grade: 'VII',  
  Hobbies: 'InternetSurfing'  
},  
{ _id: 2, StudName: 'Jannie', Grade: 'VIII', Hobbies: 'Music' },  
{ _id: 3, StudName: 'Jacob Adam', Grade: 'VII', Hobbies: 'Swimming' },  
{  
  _id: 4,  
  StudName: 'Amy Jacks',  
  Grade: 'X',  
  Hobbies: 'Dancing',  
  Location: 'Network'  
},  
{ _id: 6, StudName: 'Aryan David', Grade: 'VII', Hobbies: 'Skating' }  
]
```

```

Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Customers.insert({cust_id:1,balance:2
00,type:'S'})
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c6c71f812483cc27dd4a64') }
}
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Customers.insert({cust_id:1,balance:1
000,type:'Z'})
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c6c739812483cc27dd4a65') }
}
[Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Customers.insert({cust_id:2,balance:1
00,type:'Z'})
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c6c74d812483cc27dd4a66') }
}
[Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Customers.insert({cust_id:2,balance:1
000,type:'C'})
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c6c75e812483cc27dd4a67') }
}
[Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Customers.insert({cust_id:2,balance:5
00,type:'C'})
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c6c76e812483cc27dd4a68') }
}
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Customers.insert({cust_id:2,balance:5
0,type:'S'})
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c6c781812483cc27dd4a69') }
}
[Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Customers.insert({cust_id:3,balance:5
00,type:'Z'})
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c6c795812483cc27dd4a6a') }
}

```

```

Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.insert({RollNo:1,Age:21,Cont:t:9876,email:'antara.de9@gmail.com'});
DeprecationWarning: Collection.insert() is deprecated. Use insertOne, insertMany, or bulkWrite.
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c6c3c3812483cc27dd4a5d') }
}
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.insertOne({RollNo:1,Age:21,Cont:t:9876,email:'antara.de9@gmail.com'});
{
  acknowledged: true,
  insertedId: ObjectId('67c6c3d8812483cc27dd4a5e')
}
Atlas atlas-ws5rct-shard-0 [primary] mydb> show mydb
MongoshInvalidInputError: [COMMON-10001] 'mydb' is not a valid argument for "show".
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.insert({RollNo:2,Age:22,Cont:t:9976,email:'anushka.de@gmail.com'});
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c6c458812483cc27dd4a5f') }
}
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.insert({RollNo:3,Age:21,Cont:t:5576,email:'anubhav.de@gmail.com'});
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c6c47f812483cc27dd4a60') }
}
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.insert({RollNo:4,Age:20,Cont:t:4476,email:'pani.de9@gmail.com'});
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c6c4a2812483cc27dd4a61') }
}
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.insert({RollNo:10,Age:23,Cont:t:2276,email:'rekha.de9@gmail.com'});
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c6c4db812483cc27dd4a62') }
}
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.insertMany([{_id:3,StudName:'Jacob Adam',Grade:'VII',Hobbies:'Swimming'}, {_id:4,StudName:'Amy Jacks',Grade:'X',Hobbies:'Dancing'}])
{ acknowledged: true, insertedIds: { '0': 3, '1': 4 } }
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.insert({_id:1,StudName:'Michelle Jacintha',Grade:'VII',Hobbies:'InternetSurfing'})
{ acknowledged: true, insertedIds: { '0': 1 } }
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.insertOne({_id:2,StudName:'Janie',Grade:'VIII',Hobbies:'Music'})
{ acknowledged: true, insertedId: 2 }

```

```

Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.find().pretty()
[
  {
    _id: ObjectId('67c6c3c3812483cc27dd4a5d'),
    RollNo: 1,
    Age: 21,
    Cont: 9876,
    email: 'antara.de9@gmail.com'
  },
  {
    _id: ObjectId('67c6c3d8812483cc27dd4a5e'),
    RollNo: 1,
    Age: 21,
    Cont: 9876,
    email: 'antara.de9@gmail.com'
  },
  {
    _id: ObjectId('67c6c458812483cc27dd4a5f'),
    RollNo: 2,
    Age: 22,
    Cont: 9976,
    email: 'anushka.de@gmail.com'
  },
  {
    _id: ObjectId('67c6c47f812483cc27dd4a60'),
    RollNo: 3,
    Age: 21,
    Cont: 5576,
    email: 'anubhav.de@gmail.com'
  },
  {
    _id: ObjectId('67c6c4a2812483cc27dd4a61'),
    RollNo: 4,
    Age: 20,
    Cont: 4476,
    email: 'pani.de9@gmail.com'
  },
  {
    _id: ObjectId('67c6c4db812483cc27dd4a62'),
    RollNo: 10,
    Age: 23,
    Cont: 2276,
    email: 'rekha.de9@gmail.com'
  }
]
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.save({StudName:'Vamsi',Grade: 'VI'})

```

```
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.updateOne({_id:6,StudName:'Aryan David',Grade:'VII'},{$set:{Hobbies:'Skating'}},{upsert:true})
{
  acknowledged: true,
  insertedId: 6,
  matchedCount: 0,
  modifiedCount: 0,
  upsertedCount: 1
}
```

```
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.insert({RollNo:11,Age:22,Name :"ABC",cont:2276,email:"rea.de9@gmail.com"})
{
  acknowledged: true,
  insertedIds: { '0': ObjectId('67c6c616812483cc27dd4a63') }
}
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.update({RollNo:11,Name:"ABC"},{$set:{Name:"FEM"}})
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.updateMany({Grade:'VII'},{$set:{status:'Active'}})
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 3,
  modifiedCount: 2,
  upsertedCount: 0
}
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.updateOne({Grade:'VII'},{$set:{status:'Active'}})
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
```

```

Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.update({_id:4}, {$set:{Location:'Network'}})
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.update({_id:4}, {$unset:{Location:'Network'}})
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}
Atlas atlas-ws5rct-shard-0 [primary] mydb> db.Student.update({RollNo:10}, {$set:{email:'abhinav@gmail.com'}})
DeprecationWarning: Collection.update() is deprecated. Use updateOne, updateMany, or bulkWrite.
{
  acknowledged: true,
  insertedId: null,
  matchedCount: 1,
  modifiedCount: 1,
  upsertedCount: 0
}

```

	LAB-01	14/03/2025
		PAGE NO : DATE :
1.	Create Database in MongoDB.	
2.	use myDB; Output : Switched to db myDB	
3.	Show dbs; Output : admin config. local mydb1	
4.	Create, Read, Update, Delete.	
①	db.createCollection('student');	
	Output : ok: 1	
②	db.student.drop(); Output : true	
③	db.student.insert({id: 1, studName: "Michelle Jaentha", grade: "VII", hobbies: "Internet surfing"}) Output :-	

db.students.find()

[

{

-id: 1,

studName: 'michel joentha',

grade: 'VII',

hobbies: 'Internet surfing'

}

]

(6) update:

db.student.update({ _id: 3 }, { studName: "Anjan David", grade: "VII" })

output:

{

acknowledged: true

insertedId: 3,

matchedCount: 0

modifiedCount: 0,

updatedCount: 1

}

db.student.find()

output:

[

{

hobbies: 'Internet Surfing'

}

id : 3

grade : 'VII'

studentname : 'Anjan David'

hobbies : 'skating'

III Import data from csv file.

```
mongoimport --db student --collection admissions --type CSV --head --file /path.csv
```

IV Aggregate Function.

```
1. db.customers.aggregate({$group: {id: "$custID", totalAccBal: {$sum: "$AccBal"}}});
```

```
2. db.customers.insertMany([
```

```
  {custID: 1, AccBal: 500, Acctype: "Savings"},
```

```
  {custID: 2, AccBal: 1000, Acctype: "checkings"},
```

```
  {custID: 3, AccBal: 1500, Acctype: "savings"}])
```

Output:

[

```
  {id: 2, totalAccBal: 2000},
```

```
  {id: 1, totalAccBal: 1500},
```

```
  {id: 3, totalAccBal: 2000}
```

]

Q7 db.customers.aggregate([
 { \$match: { \$AcctType: "Savings" } },
 { \$group: { _id: "\$custID", totalAccBal: {
 \$sum: '\$AcctBal' } } }])

Output:-

[

8

{ _id: 2, totalAccBal: 1500 },
{ _id: 1, totalAccBal: 500 },
{ _id: 3, totalAccBal: 20000 }

] ✓

S.R.DR

Lab 2: Cassandra

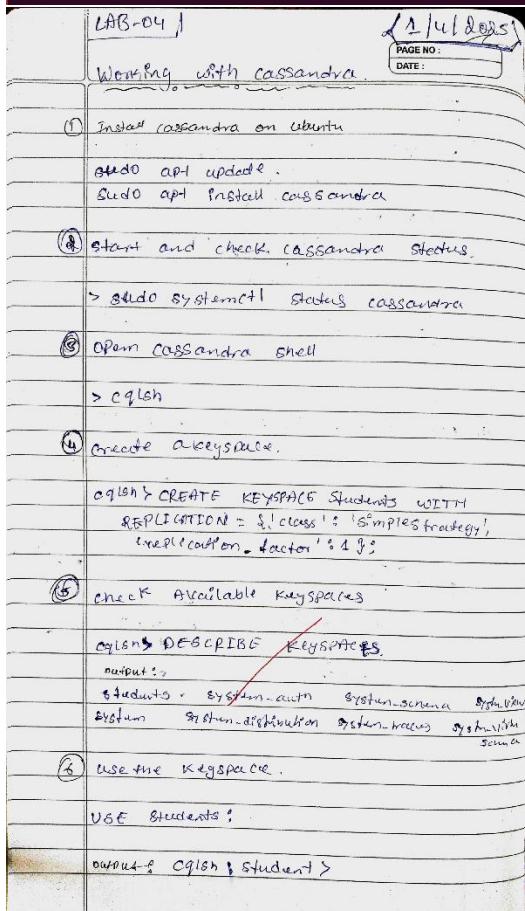
Question: Perform the following DB operations using Cassandra.

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> 
```



(7) Create a Table.

CREATE TABLE Students_Info (

Roll-No INT PRIMARY KEY,

StudentName TEXT,

DateOfJoining TIMESTAMP,

Last_Exam_Percent DECIMAL

);

Output: English : students>

(8) Insert Data

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

(2, 'Kiran', '2012-03-12', 89.9);

~~INSERT INTO Students_Info (Roll-No, StudName,~~

~~DateOfJoining, Last_Exam_Percent) VALUES~~

~~(3, 'Tarun', '2012-03-12', 78.9);~~

Output: English : students>

(9) Query All Data.

SELECT * FROM Students_Info;

Output: >

roll_no	StudName	DateOfJoining	Last_Exam_Percent
1	Asha	2012-03-12 0:0:0	79.9
2	Kiran	2012-03-12 0:0:0	89.9
3	Tarun	2012-05-12 0:0:0	78.9

PAGE NO:
DATE:

(10) Query Specific Rows

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

Output:-

roll_no	studname	dateofjoining	cgpa-exam-percentage
1	Asha	2012-03-12 0:0:0	79.9
2	Kiran	2012-03-12 0:0:0	89.9

(11) Query Using Non-indexed column

SELECT * FROM Students_Info WHERE StudName
= 'Asha';

Output:-

roll_no	studname	dateofjoining	cgpa-exam
1	Asha	2012-03-12 0:0:0	79.9

(12) Create an Index.

CREATE INDEX ON Students_Info (StudName);

(13) Update Data

UPDATE Students_Info SET StudName = 'David Sheen'
WHERE Roll_No = 2;

Output:- qqlsh & Students

(14) Update primary key.

UPDATE Students_Info SET roll_no = 6 WHERE roll_no = 3;

15) Delete value

~~DELETE students_info SET studName = 'David Sheen' WHERE rollNo = 2;~~

~~DELETE lastExamPercent FROM students_info
WHERE rollNo = 2;~~

16) Delete row

~~DELETE FROM student_info WHERE rollNo = 2;~~

17) set collection

~~ALTER TABLE student_info ADD hobbies set
(text)~~

~~Set - order of element no matters]
List : order of element matters]~~

18) UPDATE student_info

~~SET language = language + ['Hindi', 'English']
WHERE rollNo = 1;~~

~~B
NAVS~~

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	DOMS	Page No.																
<u>Cassandra:</u>																		
1. Create Keyspace Students with replication factor 'class' 3 and 'replication factor': 1}																		
2. Describe Keyspace;																		
3. Select * from system_schema.keyspaces;																		
4. use Students;																		
5. Create Table students_info (roll_no int Primary key, Studentname text, Dateofjoining timestamp, last_exam_percent double);																		
6. Begin Batch Insert into students_info (roll_no, Studentname, Dateofjoining, last_exam_percent) values (1, 'Asha', '2012-03-12', 79.9); Apply Batch;																		
7. Select * from Students_info;																		
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>roll_no</th> <th>Dateofjoining</th> <th>Last_Exam_Percent</th> <th>Studentname</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>2012-03-12</td> <td>79.9</td> <td>Asha</td> </tr> </tbody> </table>			roll_no	Dateofjoining	Last_Exam_Percent	Studentname	1	2012-03-12	79.9	Asha								
roll_no	Dateofjoining	Last_Exam_Percent	Studentname															
1	2012-03-12	79.9	Asha															
8. select * from students_info where roll_no in (1,2,3);																		
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th>roll_no</th> <th>Dateofjoining</th> <th>Last_Exam_Percent</th> <th>Studentname</th> </tr> </thead> <tbody> <tr> <td>1</td> <td></td> <td>79</td> <td>Asha</td> </tr> <tr> <td>2</td> <td></td> <td>80</td> <td>Parkha</td> </tr> <tr> <td>3</td> <td></td> <td>81</td> <td>Karishma</td> </tr> </tbody> </table>			roll_no	Dateofjoining	Last_Exam_Percent	Studentname	1		79	Asha	2		80	Parkha	3		81	Karishma
roll_no	Dateofjoining	Last_Exam_Percent	Studentname															
1		79	Asha															
2		80	Parkha															
3		81	Karishma															

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 ..//Do
wnloads/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-05 /

08/4/2025

PAGE NO:

DATE:

Perform the following DB operations using Cassandra.

① Create a Keyspace by name library

CREATE KEYSPACE library

WITH REPLICATION = {

'class' : 'SimpleStrategy'

'replication-factor' : 3 } ;

② Create a column family by name library_info with attribute Stud_Id primary key

Counter-value of type Counter,

Stud_Name , Book_Name , Book_Id ,

Date_of_Issue

CREATE TABLE library.library_info (

Stud_Id INT PRIMARY KEY

Counter-value Counter,

Stud_Name Text ,

Book_Name Text ,

Date_of_Issue TIMESTAMP

);

③ Counter table.

CREATE TABLE Book_Counter (

Stud_Id , int

Book_Name Text ,

Counter-value Counter ,

PRIMARY KEY (Stud_Id , Book_Name)

);

3) Insert the values into table in batch

BEGIN BATCH

INSERT INTO Library_Info (stud_Id, Stud_Name,
Book_Name, Book_Id, Date_of_Issue)
VALUES (112, 'John Doe', 'BDA', 'BK101',
'2025-04-01');

INSERT INTO Library_Info (stud_Id, Stud_Name,
Book_Name, Book_Id, Date_of_Issue)
VALUES (113, 'Alice Smith', 'ML', 'BK110',
'2025-04-02'),
VALUES (114, 'Charlie Brown', 'BDA', 'BK101',
'2025-04-03');

VALUES (116, 'Dairy Ray', 'DS', 'BK104',
'2025-04-06');

APPLY BATCH;

4) Display the details of the table created
& increase the value of counter.

BEGIN BATCH

UPDATE Book_Count USE, location_value ~
Count_in_value + 1 WHERE Stud_Id = 10 &
Book_Name = 'BDA';
—> 1 = 112 AND 'BDA';
—> 2 = 113 AND 'ML';
—> 3 = 114 AND 'ML';
—> 4 = 115 AND 'BDA';
—> 5 = 116 AND 'DS';

5) write a query to show that a student with id has taken a book "BOA" 2 times.

SELECT * FROM Book_Counten.

WHERE stud_id = 112 AND Book_Name = "BOA"

Output

stud_id	book_name	counten_val
112	BOA	2

6) Export the created column to a CSV file using -e "COPY library_info TO 'library_info_expert.csv' WITH HEADER =true";

SELECT * FROM library_info;

stud_id	book_id	bookname	date-of-b	stnm
114	BTK 103	A I	2025-04-03	Bobbin
113	102	ML	2025-04-02	Alicesm
116	104	DS	2025-04-06	Dairhy
112	101	BOA	2025-04-05	Johnay
115	101	BDP	2025-04-04	charit,

S
R.302

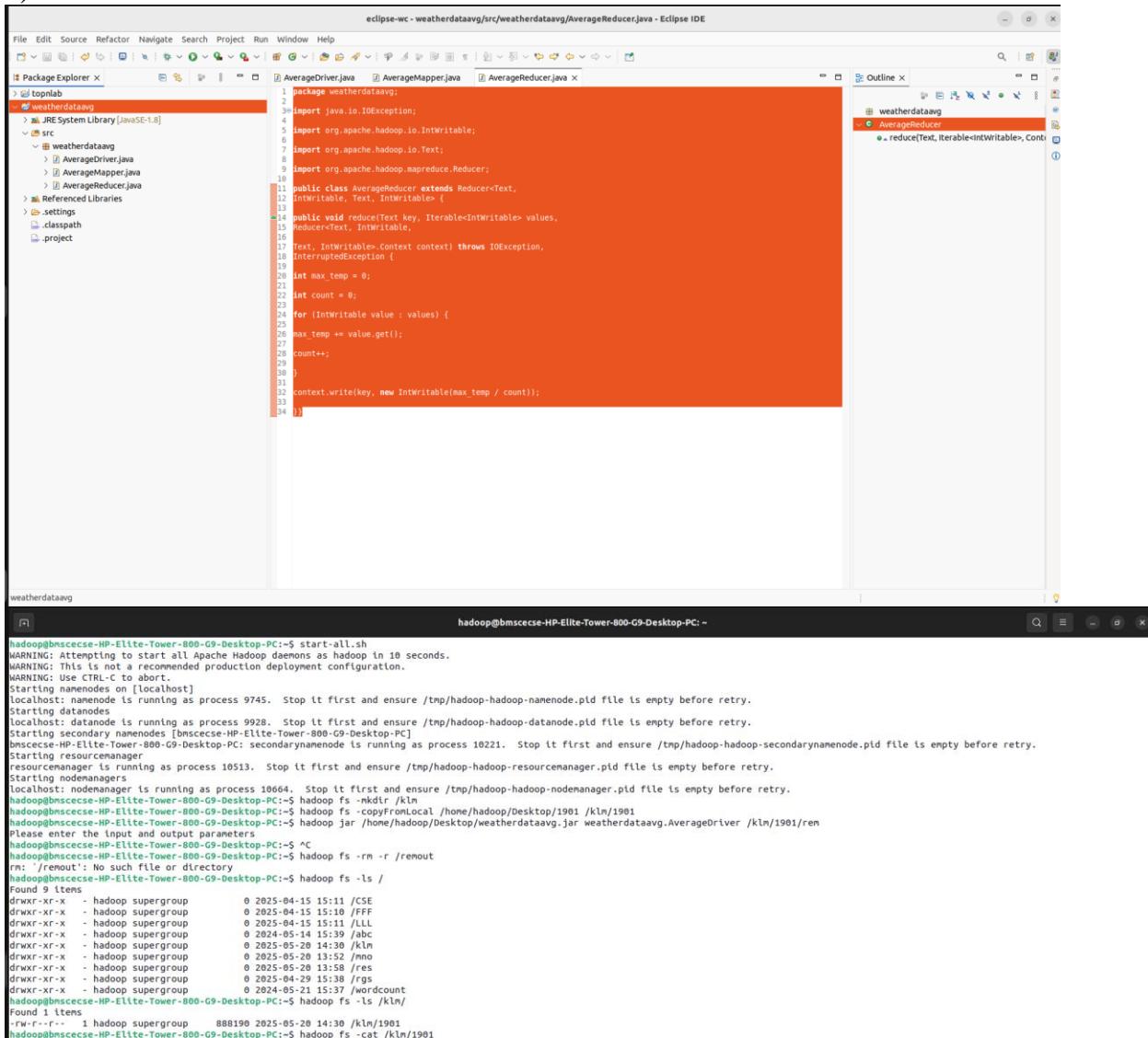
Lab 5: Hadoop

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 following details:

- Eclipse IDE Title Bar:** eclipse-wc - weatherdataavg/src/weatherdataavg/AverageReducer.java - Eclipse IDE
- Package Explorer:** Shows the project structure with packages `toplab` and `weatherdataavg`, and files `AverageDriver.java`, `AverageMapper.java`, and `AverageReducer.java`.
- Code Editor:** Displays the `AverageReducer.java` code:1 package weatherdataavg;
2
3 import java.io.IOException;
4
5 import org.apache.hadoop.io.IntWritable;
6
7 import org.apache.hadoop.mapreduce.Reducer;
8
9 import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
10 import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
11
12 public class AverageReducer extends Reducer<Text,
13 IntWritable, Text, IntWritable> {
14 public void reduce(Text key, Iterable<IntWritable> values,
15 Reducer<Text, IntWritable, Text, IntWritable>.Context context) throws IOException,
16 InterruptedException {
17 int max_temp = 0;
18 int count = 0;
19 for (IntWritable value : values) {
20 max_temp += value.get();
21 count++;
22 }
23 max_temp /= count;
24 context.write(key, new IntWritable(max_temp / count));
25 }
26 }
- Outline View:** Shows the `AverageReducer` class definition with its `reduce` method.
- Terminal Window:** Shows the command-line output of a Hadoop job run on a local host (HP-Elite-Tower-800-G9-Desktop-PC). The commands include starting daemons, creating namenodes, and running a wordcount job. The output shows file creation and deletion operations.

```

hadoop@bmseccse-HP-Elite-Tower-800-C9-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.PathOutputCommitterFactory: No output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:34:21,406 INFO output.PathOutputCommitter: 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.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.PathOutputCommitterFactory: 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.FileOutputCommitter
2025-05-20 14:34:21,440 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 kvl 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 14:34:21,580 INFO mapred.MapTask: mapred.output.collector.class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-20 14:34:21,580 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 14:34:21,595 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 14:34:21,602 INFO mapred.LocalJobRunner: 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 14:34:21,605 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 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_MAGIC=0
WRONG_REDUCE=0
File Output Format Counters
Bytes Written=8
2025-05-20 14:34:21,803 INFO mapred.LocalJobRunner: Finishing task: attempt_local261815800_0001_r_000000_0
2025-05-20 14:34:21,803 INFO mapred.LocalJobRunner: reduce task executor complete.
2025-05-20 14:34:22,000 INFO mapred.LocalJobRunner: Job attempt_local261815800_0001 running in uber mode : false
2025-05-20 14:34:22,405 INFO mapreduce.Job: map 100% reduce 100%
2025-05-20 14:34:22,407 INFO mapreduce.Job: Job job_local261815800_0001 completed successfully
2025-05-20 14:34:22,414 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=153312
FILE: Number of bytes written=1500206
FILE: Number of 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
Combine output records=0
Spilled Records=6564
Failed Shuffles=0
Merged Map outputs=0
GC time elapsed (ms)=0
Total committed heap usage (bytes)=633339904

File Input Format Counters
File Output Format Counters
2025-05-20 14:34:21,605 INFO mapred.LocalJobRunner: Finishing task: attempt_local261815800_0001_r_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 output.FileOutputCommitter: FileOutputCommitter Algorithm version is 2
2025-05-20 14:34:21,612 INFO output.FileOutputCommitter: Using ResourceCalculatorProcessTree
2025-05-20 14:34:21,612 INFO mapred.Task: Using ResourceCalculatorProcessTree
2025-05-20 14:34:21,612 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@cd4d167
2025-05-20 14:34:21,614 WARN Impl.MetricsSystemImpl: Jobtracker metrics system already initialized!
2025-05-20 14:34:21,622 INFO reduce.MergeManagerImpl: MergerManager: memoryLimit=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 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_r_000000_0 decompress: 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.LocalJobRunner: LocalJobRunner 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, 72210 bytes from disk
2025-05-20 14:34:21,645 INFO reduce.MergeManagerImpl: Merging 0 segments, 0 bytes from memory into reduce
2025-05-20 14:34:21,645 INFO mapred.Merger: Merging 1 sorted segments
2025-05-20 14:34:21,645 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 72199 bytes
2025-05-20 14:34:21,646 INFO mapred.LocalJobRunner: 1 / 1 copied.
2025-05-20 14:34:21,679 INFO Configuration.deprecation: mapred.skip.ls deprecated. Instead, use mapreduce.job.skipped
2025-05-20 14:34:21,767 INFO mapred.Task: attempt_local261815800_0001_r_000000_0 is done. And is in the process of committing
2025-05-20 14:34:21,767 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/...
2025-05-20 14:34:21,802 INFO mapred.LocalJobRunner: reduce > reduce
2025-05-20 14:34:21,802 INFO mapred.Task: Task 'attempt_local261815800_0001_r_000000_0' done.
2025-05-20 14:34:21,803 INFO mapred.Task: Final Counters for attempt_local261815800_0001_r_000000_0: Counters: 30
File System Counters
FILE: Number of bytes read=148882
FILE: Number of bytes written=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=7210
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_ID=0
CONNECTION=0
IO_ERROR=0
MOVED_PARTITION=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:

- Project Explorer:** Shows the project structure for "weatherdatameanmax" containing "MeanMaxDriver.java", "MeanMaxReducer.java", and "MeanMaxMapper.java".
- Code Editor:** Displays the "MeanMaxReducer.java" file, which is a Hadoop Reducer class. It includes imports for Java IO, org.apache.hadoop.io.IntWritable, org.apache.hadoop.io.Text, and org.apache.hadoop.mapreduce.Reducer. The code implements the reduce method to calculate the maximum temperature from a list of temperatures.
- Outline View:** Shows the class hierarchy and methods, including the reduce method in MeanMaxReducer.
- Terminal Window:** Shows a terminal session on a Linux system (hadoop@bmsecece-HP-Elite-Tower-800-G9-Desktop-PC) running Hadoop commands. The user runs "start-all.sh" and then "hadoop --help". The output provides detailed information about Hadoop's command-line options and subcommands.

```
hadoop@bmsecece-HP-Elite-Tower-800-G9-Desktop-PC:~$ start-all.sh
WARNING: Attaching to port 5432 to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
localhost: namenode is running as process 9745. Stop it first and ensure /tmp/hadoop-hadoop-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 9928. Stop it first and ensure /tmp/hadoop-hadoop-datanode.pid file is empty before retry.
Starting secondarynamenodes [bmsecece-HP-Elite-Tower-800-G9-Desktop-PC]
bmsecece-HP-Elite-Tower-800-G9-Desktop-PC: secondarynamenode is running as process 10221. Stop it first and ensure /tmp/hadoop-hadoop-secondarynamenode.pid file is empty before retry.
Starting resourcemanager
resourcemanager is running as process 10513. Stop it first and ensure /tmp/hadoop-hadoop-resourcemanager.pid file is empty before retry.
Starting nodemanagers
localhost: nodemanager is running as process 10664. Stop it first and ensure /tmp/hadoop-hadoop-nodemanager.pid file is empty before retry.
hadoop@bmsecece-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop ls -mdir /omn
ERROR: ls is not COMMAND, fully qualified CLASSNAME.
Usage: hadoop [OPTIONS] SUBCOMMAND [SUBCOMMAND ...] [OPTIONS]
Or: hadoop [OPTIONS] CLASSNAME [CLASSNAME OPTIONS]
where CLASSNAME is a user-provided Java class

OPTIONS is none or any of:

buildpaths      attempt to add class files from build tree
--config dir    Hadoop config directory
--debug         turn on shell script debug mode
--help          usage information
hostnames list[,of,host,names] hosts to use in worker mode
hosts filename  list of hosts to use in worker mode
loglevel level  set the log4j level for this command
workers         turn on worker mode

SUBCOMMAND is one of:

Admin Commands:
daemonlog      get/set the log level for each daemon

Client Commands:
archive        create a Hadoop archive
checknative    check native Hadoop and compression libraries availability
classpath      prints the class path needed to get the Hadoop jar and the required libraries
conftest       validate configuration XML files
decommission   turn off active data providers
distch        distributed metadata changes
distcp        copy file or directories recursively
dtutil        operations related to delegation tokens
envvars       display computed Hadoop environment variables
fs            run a generic filesystem user client
gridmix       submit a mix of synthetic job, modeling a profiled from production job
jar <jar>      run a jar file. NOTE: please use "yarn jar" to launch YARN applications, not this command.
minnauth     prints the Java library path
```

```

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

applications, not this command.
jnpipath prints the java.library.path
kdiag Diagnose Kerberos Problems
kerbname show auth_to_local principal conversion
key name keys via the KeyProvider
runefolder scale a given input trace
runertrace convert logs into a runen trace
siguard 53 Commands
trace view and modify Hadoop tracing settings
version print the version

Daemon Commands:

kns 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@bmsccse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mkdir /omn
hadoop@bmsccse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/weatherdatamax.jar weatherdatamax.MeanMaxDriver /omn/1901 /ren
2025-05-20 14:53:15,576 INFO Impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-05-20 14:53:15,615 INFO Impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-05-20 14:53:15,615 INFO Impl.MetricsSystemImpl: Jobtracker metrics system started
2025-05-20 14:53:15,679 WARN MapreduceJobUploader: 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: job split into splits:1
2025-05-20 14:53:15,764 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2143084439_0001
2025-05-20 14:53:15,830 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-05-20 14:53:15,891 INFO mapreduce.Job: The url to track the job: http://localhost:8880/
2025-05-20 14:53:15,891 INFO mapreduce.Job: Running job: job_local2143084439_0001
2025-05-20 14:53:15,895 INFO output.PathOutputCommitterFactory: output committer set in config null
2025-05-20 14:53:15,895 INFO output.FileOutputCommitter: output committer factory defined, defaulting to FileOutputCommitterFactory
2025-05-20 14:53:15,895 INFO output.FileOutputCommitter: FileOutputCommitter Algorithm version is 2
2025-05-20 14:53:15,895 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2025-05-20 14:53:15,896 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapred.FileOutputCommitter
2025-05-20 14:53:15,981 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-05-20 14:53:15,982 INFO mapred.LocalJobRunner: Starting task: attempt_local2143084439_0001_m_000000_0
2025-05-20 14:53:15,996 INFO output.PathOutputCommitter: 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 14:53:15,996 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:9800/omn/1901:0+888190
2025-05-20 14:53:16,044 INFO mapred.MapTask: (EQUATOR) @ kvl 26214396(104857504)
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: buffstart = 0; bufvold = 104857600
2025-05-20 14:53:16,044 INFO mapred.MapTask: bufstart = 26214396; length = 6553600
2025-05-20 14:53:16,044 INFO mapred.MapTask: bufvold = 104857600
2025-05-20 14:53:16,110 INFO mapred.LocalJobRunner:
2025-05-20 14:53:16,110 INFO mapred.MapTask: Starting flush of map output
2025-05-20 14:53:16,110 INFO mapred.MapTask: Spilling map output
2025-05-20 14:53:16,119 INFO mapred.MapTask: bufstart = 0; bufend = 45948; bufvold = 104857600
2025-05-20 14:53:16,119 INFO mapred.MapTask: bufstart = 26214396(104857504); kvend = 26188144(104752576); length = 26253/6553600
2025-05-20 14:53:16,128 INFO mapred.MapTask: Finished split 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.Task: LocalJobRunner: Done
```

```

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

INFO output.FileOutputCommitter: skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false

2025-05-20 14:53:16,140 INFO mapred.Task: Using ResourceCalculatorProcessTree

2025-05-20 14:53:16,146 INFO mapred.ReduceTask: Using ShuffleConsumerPlugin: org.apache.hadoop.mapreduce.task.reduce.Shuffle@bd64ffe0

2025-05-20 14:53:16,147 WARN Impl.MetricsSystemImpl: Jobtracker metrics system already initialized!

2025-05-20 14:53:16,155 INFO reduce.MergeManagerImpl: MergerManager: mergeThreshold=3847439360, loSortFactor=10, memToMemMergeOutputsThreshold=10

2025-05-20 14:53:16,156 INFO reduce.EventFetcher: attempt_local2143084439_0001_r_000000_0 Thread started: EventFetcher for fetching Map Completion Events

2025-05-20 14:53:16,170 INFO reduce.LocalFetcher: localfetcher#1 about to shuffle output of map_attempt_local2143084439_0001_m_000000_0 decomp: 59078 len: 59082 to MEMORY

2025-05-20 14:53:16,172 INFO reduce.InMemoryMapOutput: Read 59078 bytes from map-output for attempt_local2143084439_0001_m_000000_0

2025-05-20 14:53:16,172 INFO reduce.LocalFetcher: cleanupTemporaryFile -> map-output of size: 59078, lnMemoryMapOutputs.size() -> 1, commitMemory -> 0, usedMemory -> 59078

2025-05-20 14:53:16,173 INFO reduce.EventFetcher: EventFetcher is interrupted.. Returning

2025-05-20 14:53:16,174 INFO mapred.LocalJobRunner: 1 / 1 copied

2025-05-20 14:53:16,174 INFO reduce.MergeManagerImpl: finalMerge called with 1 in-memory map-outputs and 0 on-disk map-outputs

2025-05-20 14:53:16,176 INFO mapred.Merger: Merging 1 sorted segments

2025-05-20 14:53:16,176 INFO mapred.Merger: Down to the last merge-pass, with 1 segments left of total size: 59073 bytes

2025-05-20 14:53:16,180 INFO reduce.MergeManagerImpl: Merged 1 segments, 59078 bytes to disk to satisfy reduce memory limit

2025-05-20 14:53:16,181 INFO reduce.MergeManagerImpl: 1 file, 59082 bytes from disk

2025-05-20 14:53:16,181 INFO mapred.Merger: 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 14:53:16,182 INFO configuration.deprecation: mapred.skip.on.ls deprecated. Instead, use mapreduce.job.skiprecords

2025-05-20 14:53:16,273 INFO mapred.Task: Task:attempt_local2143084439_0001_r_000000_0 is done. And is in the process of committing

2025-05-20 14:53:16,273 INFO mapred.Task: LocalJobRunner: Done

2025-05-20 14:53:16,273 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: Task: Final Counters for attempt_local2143084439_0001_r_000000_0: Counters: 30

File System Counters

- FILE: Number of bytes read=122769
- FILE: Number of bytes written=763193
- FILE: Number of read operations=0
- FILE: Number of large read operations=0
- FILE: Number of write operations=0
- HDFS: Number of bytes read=888190
- HDFS: Number of bytes written=81
- HDFS: Number of read operations=10
- HDFS: Number of large read operations=0
- HDFS: Number of write operations=3
- HDFS: Number of bytes read erasure-coded=0

Map-Reduce Framework

- Combine input records=0
- Combine output records=0
- Reduce input groups=12
- Reduce shuffle bytes=100882
- Reduce input records=6564
- Reduce output records=12
- Spilled Records=6564
- Shuffled Maps =1
- Failed Shuffles=0
- Merged Map outputs=1
- GC time elapsed (ms)=0
- Total committed heap usage (bytes)=526305152

Shuffle Errors

- BAD_ID#
- CONNECTION-A

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+88190
2025-05-20 14:53:16,044 INFO mapred.MapTask: EQUATOR: 0 kv: 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 14:53:16,110 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 14:53:16,128 INFO mapred.MapTask: Finished spilt 0
2025-05-20 14:53:16,135 INFO mapred.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.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
  File System Counters
    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
    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=1770988
    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 hwm=0

```

Screenshot captured
You can paste the image from the clipboard.

```

hadoop@bmscsece-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
  Redundant splits=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_TYPE=0
  WRONG_DURUCE=0
File Input Format Counters
  Bytes Read=888190
File Output Format Counters
  Bytes Written=81
hadoop@bmscsece-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@bmscsece-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@bmscsece-HP-Elite-Tower-800-G9-Desktop-PC: ~

```

Hadoop Exercise

→ Start Hadoop.

\$ start -cell.sh

→ Create directory

hdfs dfs mkdir /bdar/hadoop

→ List contents

hadoop fs - ls l

→ put - copy from Local to HDFS

hdfs dfs -copy from Local /home/Desktop/
bdar/local.txt /bdar/hadoop/file.txt

→ cat - display file contents.

hdfs dfs -cat /bdar/hadoop/file.txt
"Hello, BASIC!"hdfs dfs -cat /bdar/hadoop/file.txt
get (download from HDFS)hdfs dfs -get /bdar/hadoop/file.txt
/home/Desktop/downloaded-file.txt

→ get merge(multiple files)

hdfs dfs -get merge% /bdar/hadoop/
file.txt

→ get fact

hadoop fs -get fact : (bdar_hadoop)

→ copy to local.

hdfs dfs -copy to local /abc/wc-tat
(home/Desktop)

→ mv (move file / directory)

hadoop fs -mv /bdar_hadoop /abc

get fact output :

file ? (bdar_hadoop)

owner ? hadoop

group ? supergroup

user :: root

group :: root

other :: root

→ CP (copies a file from one directory
to another directory)

hadoop fs -cp /home/Desktop/bdar-
hadoop

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 project structure under "word_count". It includes a JRE System Library [JavaSE-1.8], a src folder containing WCDriver.java, WCMapper.java, and WCReducer.java, and a wordcount folder.
- WCDriver.java Content:**

```
1 import java.io.IOException;
2 import org.apache.hadoop.conf.Configured;
3 import org.apache.hadoop.fs.Path;
4 import org.apache.hadoop.io.IntWritable;
5 import org.apache.hadoop.io.Text;
6 import org.apache.hadoop.mapred.FileInputFormat;
7 import org.apache.hadoop.mapred.FileOutputFormat;
8 import org.apache.hadoop.mapred.JobClient;
9 import org.apache.hadoop.mapred.JobConf;
10 import org.apache.hadoop.util.Tool;
11 import org.apache.hadoop.util.ToolRunner;
12
13 public class WCDriver extends Configured implements Tool {
14
15     public int run(String[] args) throws IOException {
16         if (args.length < 2) {
17             System.out.println("Please provide input and output paths");
18             return -1;
19         }
20
21         JobConf conf = new JobConf(WCDriver.class);
22         conf.setJobName("WordCount");
23         conf.setJarByClass(WCDriver.class); // Ensures job runs from correct JAR
24
25         FileInputFormat.setInputPaths(conf, new Path(args[0]));
26         FileOutputFormat.setOutputPath(conf, new Path(args[1]));
27
28         conf.setMapperClass(WCMapper.class);
29         conf.setReducerClass(WCReducer.class);
30
31         conf.setMapOutputKeyClass(Text.class);
32 }
```
- Task List:** Shows a single task named "WCDriver".
- Outline:** Shows the outline of the WCDriver class.
- Problems:** Shows a message about Java 24 being released and options to install support or download a JDK.

```

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@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/wordcount.jar WCDriver /rsg/sample.txt /result
2025-05-06 15:05:01,299 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 kvi 26214396(104857584)
2025-05-06 15:05:01,671 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-05-06 15:05:01,671 INFO mapred.MapTask: soft limit at 83886080
2025-05-06 15:05:01,671 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-05-06 15:05:01,671 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-05-06 15:05:01,673 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-05-06 15:05:01,742 INFO mapred.LocalJobRunner:
2025-05-06 15:05:01,742 INFO mapred.MapTask: Starting flush of map output
2025-05-06 15:05:01,742 INFO mapred.MapTask: Spilling map output
2025-05-06 15:05:01,742 INFO mapred.MapTask: bufstart = 0; bufend = 169; bufvoid = 104857600
2025-05-06 15:05:01,742 INFO mapred.MapTask: kvstart = 26214396(104857584); kvend = 26214320(104857280); length = 77/6553600
2025-05-06 15:05:01,745 INFO mapred.MapTask: Finished spill 0
2025-05-06 15:05:01,751 INFO mapred.Task: Task:attempt_local90897529_0001_m_000000_0 is done. And is in the process of committing
2025-05-06 15:05:01,753 INFO mapred.LocalJobRunner: hdfs://localhost:9000/rsg/sample.txt:0+89
2025-05-06 15:05:01,753 INFO mapred.Task: Task 'attempt_local90897529_0001_m_000000_0' done.
2025-05-06 15:05:01,756 INFO mapred.Task: Final Counters for attempt_local90897529_0001_m_000000_0: Counters: 23
    File System Counters
        FILE: Number of bytes read=4273
        FILE: Number of bytes written=639534
        FILE: Number of read operations=0

```

```

hadoop@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@bmscse-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 15:05:01,750 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, toSortFactor=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@bmscsecse-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 mapred.Merger: Merging 0 segments, 0 bytes from memory into reduce
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@bmscsecse-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@bmscsecse-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@bmscsecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ 
```

a) Find average temperature for each year from NCDC dataset.

Average temp;

```
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
public class AverageDriver {
    public static void main(String[] args) throws Exception {
        if (args.length > 0) {
            System.out.println("please enter input & output parameters");
            System.exit(1);
        }
        Job job = new Job();
        job.setJarByClass(AverageDriver.class);
        job.setJobName("max temperature");
        FileInputFormat.addInputPath(job, new Path(args[0]));
        FileOutputFormat.setOutputFormat(job,
            new Path(args[1]));
        job.setMapperClass(AverageMapper.class);
        job.setReducerClass(AverageMap);
    }
}
```

SP22
Batch

Week 7 continues

29/4/2023

PAGE NO:

DATE:

Reducer Code :-

```
/* import all the required library */
```

```
public class WCReducer extends mappredu...
```

Base implements Reducer<Text, Implement...

```
Text, IntWritable> {
```

```
public void reduce(Text key, Iterator<IntWrit...
```

able> value, OutputCollector<Text, IntWrit...

```
able> output,
```

Reporter reporter) throws IOException

8

```
int count = 0;
```

```
while (value.hasNext()) {
```

```
IntWritable i = value.next();
```

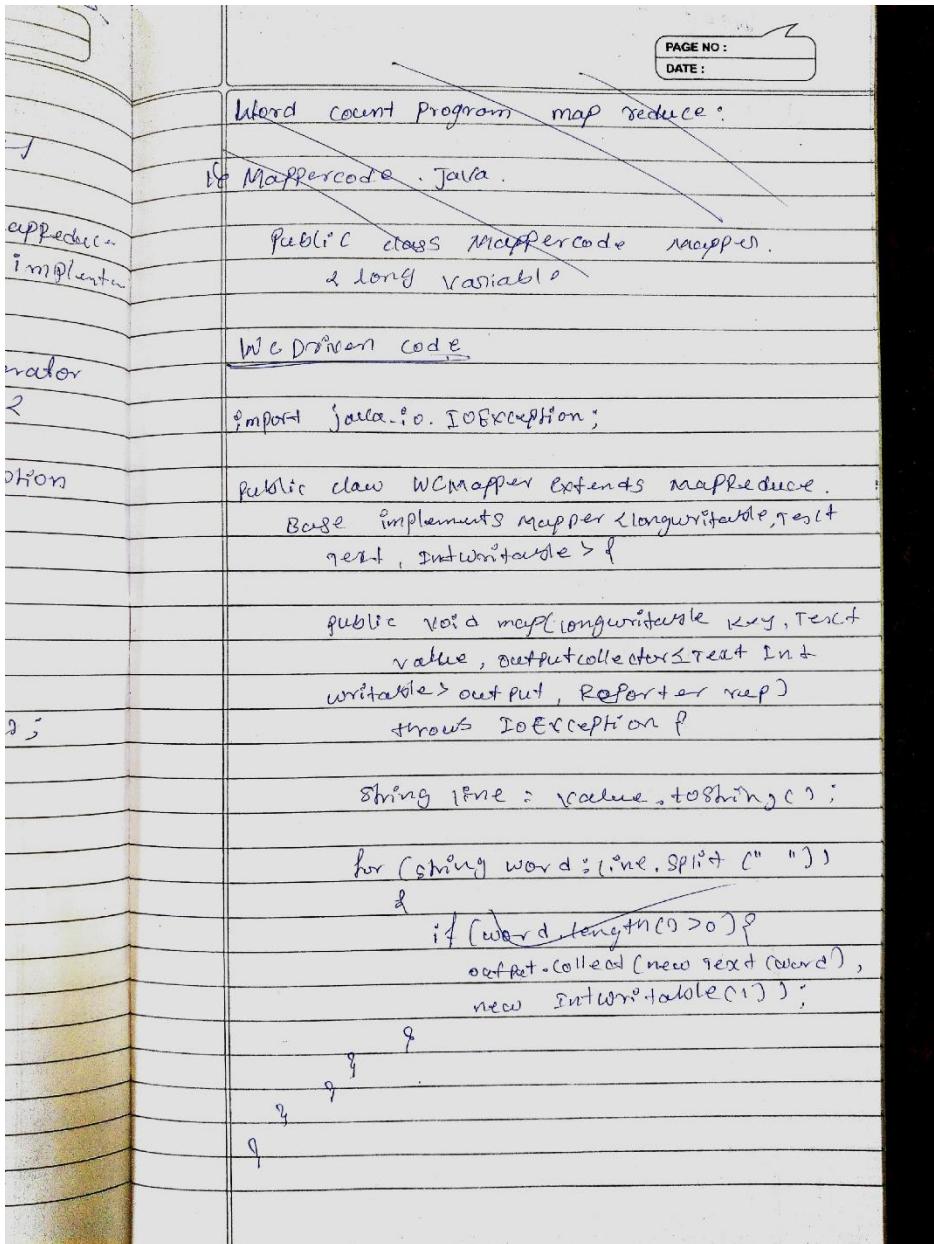
```
count += i.get();
```

9
9

```
output.collect(key, new
```

```
IntWritable(count));
```

S.R. &



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:

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

- Toolbar:** File, Edit, Source, Refactor, Navigate, Search, Project, Run, Window, Help.
- Package Explorer:** Shows the project structure with `topnlab` as the root, containing `JRE System Library [JavaSE-1.8]`, `src`, and `TopNMapper.java`.
- Editor:** Displays the `TopNMapper.java` file content. The code implements the `Mapper<Object, Text, IntWritable>` interface. It includes imports for `java.io.IOException`, `Mapper`, `Object`, `Text`, `IntWritable`, and `Context`. The `map` method tokenizes input lines, removes punctuation, converts them to lowercase, and emits words with their counts.
- Outline View:** Shows the class structure with methods `map` and `reduce`, and fields `one`, `word`, and `tokens`.

```
File Edit Source Refactor Navigate Search Project Run Window Help
eclipse-wc - topnlab/src/topnlab/TopNMapper.java - Eclipse IDE

File Edit Source Refactor Navigate Project Run Window Help
Package Explorer X
JRE System Library [JavaSE-1.8]
src
  topnlab
    TopN.java
    TopNReducer.java
    TopNMapper.java X
    TopNCombiner.java
  settings
  .classpath
  .project

TopNMapper.java
1 package topnlab;
2
3 import java.io.IOException;
4
5 public class TopNMapper extends Mapper<Object, Text, IntWritable> {
6
7     private static final IntWritable one = new IntWritable(1);
8     private Text word = new Text();
9
10    // Regex to clean out punctuation and special symbols
11    private String tokens = "[\\$#>\\\\^\\\\{\\\\}\\\\\\\\\\\\;,\\\\:-\\\\?\\\\!\\\\*\\\\]";
12
13    @Override
14    public void map(Object key, Text value, Mapper<Object, Text, IntWritable>.Context context)
15        throws IOException, InterruptedException {
16
17        // Clean line and convert to lowercase
18        String cleanLine = value.toString().toLowerCase().replaceAll(this.tokens, " ");
19
20        // Tokenize by whitespace
21        StringTokenizer itr = new StringTokenizer(cleanLine);
22
23        // Emit each word with a count of 1
24        while (itr.hasMoreTokens()) {
25            this.word.set(itr.nextToken().trim());
26            context.write(this.word, one);
27        }
28    }
29}
30
31
32
33
34
35

Outline X
topnlab
  TopNMapper
    <one : IntWritable>
      word : Text
      tokens : String
    <map : Object, Text, Context> : void
```

```

hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ start -all.sh
Command 'start' not found, did you mean:
  command 'stars' from snap stars (2.7jrc3)
  command 'rstart' from deb x11-session-utils (7.7+4build2)
  command 'kstart' from deb kde-cli-tools (4:5.24.4-0ubuntu1)
  command 'startx' from deb xinit (1.4.1-0ubuntu4)
  command 'stat' from deb coreutils (8.32-4.1ubuntu1.2)
  command 'tart' from deb tart (3.10-1build1)
See 'snap info <snapname>' for additional versions.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ start-all.sh
WARNING: Attempting to start all Apache Hadoop daemons as hadoop in 10 seconds.
WARNING: This is not a recommended production deployment configuration.
WARNING: Use CTRL-C to abort.
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [bmscecse-HP-Elite-Tower-800-G9-Desktop-PC]
Starting resourcemanager
Starting nodemanagers
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mkdir /mno
mkdir: Cannot create directory /mno. Name node is in safe mode.
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ ^C
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ ^C
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hdfs dfsadmin -safemode get
Safe mode is OFF
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -mkdir /mno
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop fs -copyFromLocal /home/hadoop/Desktop/sample.txt /mno/sample.txt
t
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/topn.jar TopN /mno/sample.txt /res
Exception in thread "main" java.lang.ClassNotFoundException: TopN
  at java.base/java.net.URLClassLoader.findClass(URLClassLoader.java:476)
  at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:594)
  at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:527)
  at java.base/java.lang.Class.forName0(Native Method)
  at java.base/java.lang.Class.forName(Class.java:398)
  at org.apache.hadoop.util.RunJar.run(RunJar.java:321)
  at org.apache.hadoop.util.RunJar.main(RunJar.java:241)
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/topn.jar topnlab /mno/sample.txt /res
Exception in thread "main" java.lang.ClassNotFoundException: topnlab
  at java.base/java.net.URLClassLoader.findClass(URLClassLoader.java:476)
  at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:594)
  at java.base/java.lang.ClassLoader.loadClass(ClassLoader.java:527)
  at java.base/java.lang.Class.forName0(Native Method)
  at java.base/java.lang.Class.forName(Class.java:398)
  at org.apache.hadoop.util.RunJar.run(RunJar.java:321)
  at org.apache.hadoop.util.RunJar.main(RunJar.java:241)
hadoop@bmscecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ hadoop jar /home/hadoop/Desktop/topn.jar TopN /mno/sample.txt /res
Exception in thread "main" java.lang.NoClassDefFoundError: topnlab/TopN (wrong name: TopN)
  at java.base/java.lang.ClassLoader.defineClass1(Native Method)

```

```
hadoop@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=14573633
28, mergeThreshold=3847439360, ioSortFactor=10, memToMemMergeOutputsThreshold=10
2025-05-20 13:58:10,038 INFO reduce.EventFetcher: attempt_local408680812_0001_r_000000_0 Thread started: EventFetcher for 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:~$
```

Week 8

6/5/05

PAGE NO :

DATE :

Scala :-

Word count :-

Object word count {

```
def Main(args: Array[String]): Unit = {  
    val input = Scala.io.StdIn.readLine("Enter  
    text (press enter to use default): ")  
    val text = if (input.trim.isEmpty) "This  
    is a test" else input.  
    val wordCount1 = countWords(text)  
    println(s"Word count : $wordCount1")  
}
```

```
def countWords(s: String): Int = {
```

```
    s.trim.split(" \\S+").count(_ nonEmpty)
```

Output:-

Enter text (press enter to use default):

Regular relational databases cannot handle linear
scaling, seamless data distribution, and other
big data requirements as efficiently as
cassandra.

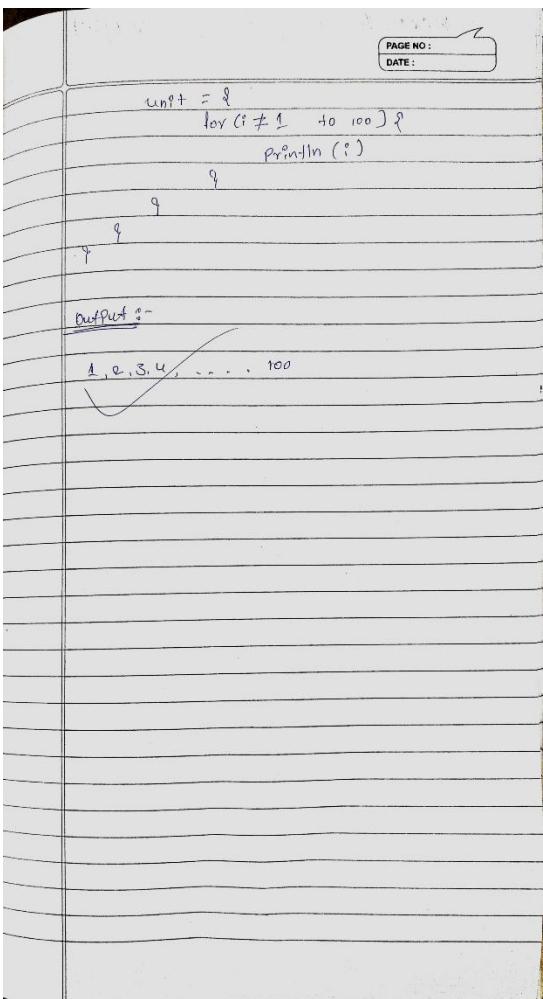
Word Count : 34

⑧

Print numbers

Object PrintNumbers {

```
    def main (args : Array[String]):
```



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
```

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

The terminal output shows the execution of the Scala program. The code prints numbers from 1 to 100 using a for loop. The nano editor screenshot shows the source code for the program.

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:

```
bsmsecse@bsmsecse-HP-Elite-Tower-800-G9-Desktop-PC:~$ spark-shell
25/05/20 15:32:38 WARN Utils: Your hostname, bsmsecse-HP-Elite-Tower-800-G9-Desktop-PC resolves to a loopback address: 127.0.1.1
; using 10.124.2.8 instead (on interface eno1)
25/05/20 15:32:38 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark/jars/spark-unsafe_2.12-3.0.3.jar) to 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

    / \ / \
   / \ \ - \ \ - / \ / \
  / \ / . \ \ / \ / \ / \
 / \ / \ / \ / \ / \ / \ / \
version 3.0.3

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

scala> val textFile = sc.textFile("/home/bsmsecse/Desktop/sparkdata.txt")
textFile: org.apache.spark.rdd.RDD[String] = /home/bsmsecse/Desktop/sparkdata.txt MapPartitionsRDD[1] at textFile at <console>:24
4

scala>

scala> val counts = textFile
counts: org.apache.spark.rdd.RDD[String] = /home/bsmsecse/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(_.value > _.value): _*)
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>

```

Week 9)

12/15/2022
PAGE NO:
DATE:

(1) Using RDD and flatmap count how many times each word appears in file.

```

from pyspark import SparkContext
sc = SparkContext(appName = "wordcount")

filePath = "path + read"
textFile = sc.textFile(filePath)
words = textFile.flatMap(lambda line: line.split())
wordPairs = words.map(lambda word: (word, 1))
result = wordPairs.collect()
for word, count in result:
    print(f'{word}: {count}')

sc.stop()

```

Output:

```

textFile = sc.parallelize([
    "hello world",
    "hello spark",
    "hello helloworld spark spark"
])

([('hello', 4), ('spark', 4)])

```

Q) Java Hadoop Mapreduce Program: Top 10 frequent words in alphabetical order.

Mapper class

```
import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
```

```
public class WordCountMapper extends Mapper<LongWritable,
    Text, Text, IntWritable> {
```

```
private final static IntWritable one = new IntWritable(1);
```

```
private Text word = new Text();
```

```
public void map(LongWritable key, Text value, Context context)
    throws IOException, InterruptedException {
```

```
String[] words = value.toString().toLowerCase().split("\\w+");
```

```
for (String w : words) {
```

```
if (!w.isEmpty()) {
```

```
word.set(w);
```

```
context.write(word, one);
```

```
}
```

```
}
```

```
}
```

Reducer class

```
import java.io.IOException;
import java.util.*;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;
```

Public class ToplineWords {

 Public static void main(String[] args) {
 throws Exception {

 Configuration conf = new Configuration();

 Job job = job.getInstance(conf, "topline");

 job.setMapperClass(WordCount.class);

 job.setMapOutputValueClass(IntWritable.class);

 job.setOutputKeyClass(Text.class);

 System.exit(job.waitForCompletion(true)); 0: 1);

}
{

Output:-

apple banana apple mango banana orange

apple apple banana grape mango apple

banana grape.

apple 5

banana 4

grape 2

mango 8

orange 1

③ Cleaning the text:

from pyspark import SparkContext

import re

```
def clean_text(rdd):
    clean_rdd = rdd.map(lambda x: x.strip())
    return clean_rdd

def main():
    sc = SparkContext(appName="textclean")
    ssc = StreamingContext(sc, 1)
    port = 999
    lines = ssc.socketTextStream(port)
    cleaned_lines = clean_text(lines)
    cleaned_lines.pprint()
    ssc.start()
    ssc.awaitTermination()

if name == "main":
    main()
```

Output :-

This is an example of Streaming text
Spark is very useful for Big Data processing

Example Streaming text

S. Raja
2015/16

Spark useful big data processing.

```
public class WordCountReducer extends Reducer<
    Text, IntWritable, Text, IntWritable> {
    private Map<String, Integer> countMap = new
        HashMap<String, Integer>();
```

```
    public void reduce(Text key, Iterable<IntWritable>
        values, Context context) {
        int sum = 0;
        for (IntWritable val : values) {
            sum += val.get();
        }
        countMap.put(key.toString(), sum);
    }
}
```

```
protected void cleanup(Context context) throws
    IOException, InterruptedException {
    countMap.entrySet().stream()
        .sorted(Map.Entry.comparingByValue(Comparator
            .reversingOrder()))
        .thenComparing(Map.Entry.comparingByKey())
}
```

```
.limit(10)
```

```
.forEach(Entry -> {

```

```
    try {

```

```
        context.write(new Text(Entry.getKey()),
            new IntWritable(Entry.getValue()));
    } catch (IOException | InterruptedException e) {

```

```
        e.printStackTrace();
    }
}
```

```
});
```

```
}
```